CHIP/PACKAGE CO-DESIGN METHODOLOGIES FOR RELIABLE 3D ICS

A Dissertation
Presented to
The Academic Faculty

by

Taigon Song

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2015

CHIP/PACKAGE CO-DESIGN METHODOLOGIES FOR RELIABLE 3D ICS

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Suresh K. Sitaraman
School of Mechanical Engineering
Georgia Institute of Technology

Dr. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: October 22, 2015

To my wife Keeon Jung and our baby,

my parents,

and my family.

ACKNOWLEDGEMENTS

Finishing my Ph.D. has been a journey with many surprises and challenges. It would not have been possible without many help. Everyone deserves proper appreciation, and I would like to use this section to thank all of those who helped me along the way.

Firstly, I would like to thank my advisor, Dr. Sung Kyu Lim, for guiding me and shaping my research. Changing my specialty from one to another was not easy, and without his guidance I would not have successfully made it. Thanks to him, my chance of pursuing the highest academic degree possible was a great success in one of the best schools in the world. I would like to thank Dr. Arijit Raychowdhury and Dr. Sudhakar Yalamanchili for guidance and suggestions on my research. In addition, I thank Dr. Saibal Mukhopadhyay and Dr. Suresh K. Sitaraman for their time and for the serving on my dissertation defense committee.

I thank Synopsys StarRC R&D group for the technical resources, discussion, and guidance regarding my face-to-face bonding research: Arthur Nieuwoudt, Ralph Iverson, and Alexander Mirgorodskiy. I also thank Beifang Qiu and Baribrata Biswas for giving me a wonderful opportunity to work at Synopsys Inc. In addition, I also thank Vivek Mishra, who was interning in Synopsys at that time with me, at the University of Minnesota for valuable discussions as well. I thank GTCAD members, Dr. Daehyun Kim, Dr. Xin Zhao, Dr. Young-Joon Lee, Dr. Krit Athikulwongse, Dr. Moongon Jung, Dr. Shreepad Panth, Chang Liu, Neela Lohith, Mohit Pathak, Yarui Peng, Sandeep Samal, Kyungwook Chang, Bon Woong Ku, Kartik Acharya, Kwangmin Kim, and Yoo-Jin Chae for their valuable comments and feedback. I would also not want to forget Pamela Halverson for her awesome help in all aspects of my office days for the last years.

I wish to thank my friends and roommates who supported me all the time: Dr. Sangkil

Kim, Jeewoong Kim, Junho Lee, Jonha Lee, Youngchul Park, Danny Kang, Yongmin Choi, Hyun Choi, and Sanggyu min. My church mentors and friends were also a big help to me: Joanna Jung, HK Bahk, Stella Song, Joohwan Kim, Michael Choi, Rachael Park, Michael Nam, Justin Cho, Letta Lee, Jinoh Yun, Mindy Yun, and Dongsik Chang. I would also like to extend my special thanks to Seung-ho Ok, Younsik Park, Daehyun Kim (my senior from Yonsei Univ.), and my mentor Jungwon Bae, who have always motivated me and helped me during my Ph.D. I thank my band, "Loveless" (Wonchul Kim, Juyoung Lee, Minjae Kang, Minkyung Hwang, Harum Chang, Eunjung Noh, Seungkyu Chang, Hwisuk Yang, Jiyoon Choi, Jihye Seo, Byunghwa Kim, and Jimin Park), and my korean friends: Shawn Kim, Junyoung Jung, Eden Joo, Chan-Ho Kong, Ki-Joon Kim, and Seung-Ho Choi, who waited me in my home country, I thank them as well.

I am particularly thankful to wife, Keeon Jung, our little baby, and my parents, who have always been on my side, for their love and encouragement throughout my life. I also thank my sister, Hyunyoung Song, for the love and support. Last but not least, I thank God and all the professors, teachers, families, and friends who guided me to become the person that I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The objective of this research is to study and develop computer-aided-design (CAD) methodologies for reliability in chip-package co-designed three-dimensional integrated circuit (3D IC) systems. 3D IC technologies refer to many vertical integration methodologies (such as through-silicon vias and face-to-face bumps) that enable the stacking of ICs. By 3D IC stacking, various benefits in terms of power and performance can be gained. However, it is not only the 3D IC design itself but also the design of the package and its many connections that must be optimized to maximize the benefit of 3D IC technology. Therefore, this work proposes design methodologies that enable reliable 3D IC in terms of signal integrity, power integrity, and thermal optimization.

The first section of this dissertation presents chip/package/PCB co-analysis methodologies. In detail, two studies are presented: (1) a methodology of co-simulating IR-drop noise for 3D IC, silicon interposer, and PCB simultaneously, and (2) a thermal analysis methodology on integrated voltage regulators (IVRs) that are implemented in silicon interposers. By proposing co-analysis methodologies in two different domains, this section provides ideas for co-analysis that can be further extended to other analysis domains as well.

The second section investigates the impact of electric coupling between through-silicon vias (TSVs) in 3D ICs. TSV-to-TSV coupling is non-negligible, and the impact of coupling is different in ICs and interposer/package/PCBs. Therefore, the first part of this section investigates how TSV-to-TSV coupling is different in ICs compared to interposers/packages and PCBs. Then, the second part proposes a methodology of analyzing TSV-to-TSV coupling in full-chip scale.

The third section investigates the impact of parasitics in face-to-face (F2F) bonding. As technology scales in F2F bonded 3D ICs, the distance between the ICs becomes as small

as few microns. Due to this shorter distance, significant electric coupling occurs between these ICs. The impact of parasitics in F2F bonding in terms of capacitance is first investigated in various scenarios. Then, a holistic methodology of extracting F2F capacitance is proposed in full-chip scale. Based on the methodology, impact of F2F parasitics in timing and power are observed.

The final section presents power reduction methodologies and its benefits when 3-tier 3D ICs are designed in OpenSPARC T2 benchmark. It is shown that one additional tier available in 3-tier 3D ICs does offer more power saving compared with their 2-tier 3D IC counterparts, but more careful floorplanning, through-silicon via (TSV) management, and block folding considerations are required. This section develops effective CAD solutions that are seamlessly integrated into commercial CAD tools to handle 3-tier 3D IC power optimization under various bonding style options.

CHAPTER I

INTRODUCTION

For the last fifty years, the semiconductor industry has been driven to double the number of transistors every two years by the "Moore's law", which motivates power/performance improvement by device scaling. This law has been used to set targets for research and development in the semiconductor industry to guide long-term planning [15]. Thanks to the Moore's law, in addition to the exponential growth in the transistor numbers by scaling, significant improvement has also been made to the performance of the transistors themselves. The development of new technologies and devices such as strained silicon [77], high-K metal gate [57], finFETs [6], and fully-depleted SOI [20] are some examples of the research done to follow this technology trend. However, doubts are rising that Moore's law may come to an end in the near future.

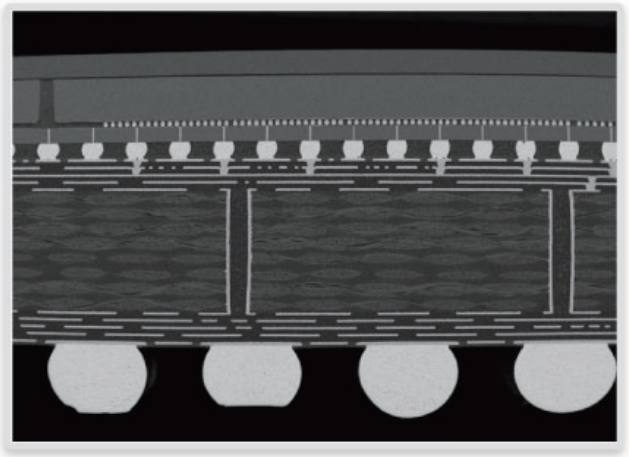
Recent studies are reporting challenges to the semiconductor scaling. First, studies indicate the physical limit of scaling. Current 14nm node transistors consist of countable number of atoms. Knowing that transistors would not be smaller than a few or less atoms, studies are predicting that scaling of transistors will eventually come to an end. Second, mask lithography is encountering its challenges. Mask lithography is currently based on 193nm lithography tools. Many technologies have been developed to extend the use of 193nm waves such as double patterning [17] and triple patterning [19]. However, a next generation lithography technology is required to follow up the mask generation in the scaling trend, and extreme ultraviolet (EUV) is a rising technology to break the lithography wall. Unfortunately, studies are still in progress to provide EUV for mass production and it suggests that EUV will come in long effort with high cost [47].

1.1 Three-dimensional Integrated Circuits (3D ICs) and Silicon Interposers as Alternative Technologies

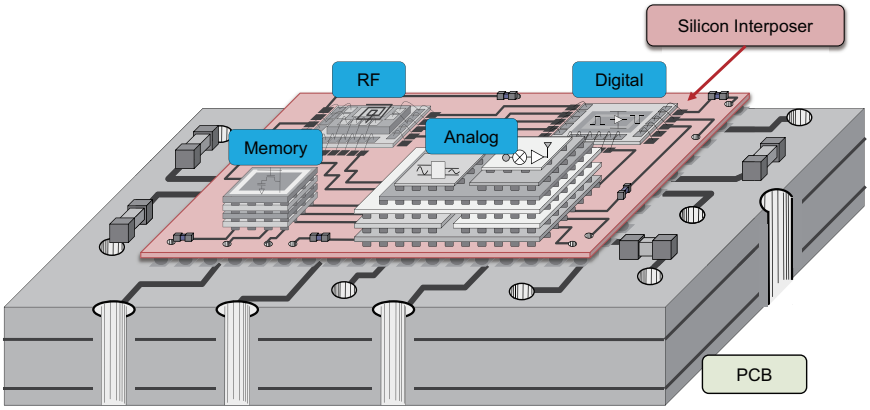
Knowing that the forecasts on semiconductor scaling is not that bright, alternative technologies for scaling are rising up. Nanowire transistors are gaining attention as a future device to replace CMOS [22], and carbon-nanotube field-effect transistors also show its potential as an alternative to CMOS based on its 20x power-performance benefits [60]. In addition to these devices, Silicon interposers and three-dimensional integrated circuits (3D ICs) are gaining significant attention as alternative technologies.

Silicon interposer, a silicon die with no actives, is a technology developed to fill the gap between ICs and packages due to the smaller interconnect it can provide in low cost. Smaller interconnects in silicon interposers allow ICs to be placed side-by-side. Thus, high-bandwidth and low-latency designs are possible. In addition, products are already made in silicon interposer [81] proving the potential of this new technology [see Figure 1 (a)]. 3D IC is a technology of stacking two (or more) ICs in vertical (3D) dimension. Comparing to conventional 2D ICs, 3D ICs provide smaller footprint because we have multiple layers of transistors instead of one. Having the smaller footprint advantage, 3D ICs can be designed to provide higher performance on lower power. As in Figure 1 (b), future roadmap of 3D ICs and silicon interposers predict that these two technologies will be combined together for ultra-miniaturized high-performance and low-power systems. This will combine every electronic components such as digital, analog, RF, and memory in a small footprint for future systems such as mobile applications [2].

3D ICs can be bonded in two different bonding styles to realize the high-performance and low-power benefits: Using through-silicon-vias (TSVs) or using face-to-face (F2F) bumps. TSVs are metal pillars that penetrate through the silicon substrate. For 3D ICs that use TSVs, ICs are bonded using the back side (where the TSV is exposed) of one die and the face side (the side where top-metal is exposed) of another die. However, in F2F, the ICs are bonded by using both face sides as the bonding side using F2F bumps. Several studies



(a)



(b)

Figure 1: (a) Silicon interposer in actual product [81] and (b) illustration of 3D ICs and silicon interposers for future ultra-miniaturized systems.

indicates that F2F 3D ICs provide advantages over TSV-based 3D ICs in many applications since they do not use any silicon area [27]. F2F bonding can also be applied by using direct copper-to-copper (Cu-Cu) bonding [62]. When Cu-Cu bonding is applied, F2F dies do not have any space between them.

1.2 Challenges

Despite the advantages 3D ICs and silicon interposers can provide, many technical challenges exist in its manufacturing and design. In the manufacturing side, for example, manufacturing reliable TSVs is very important. However, faults during manufacturing such

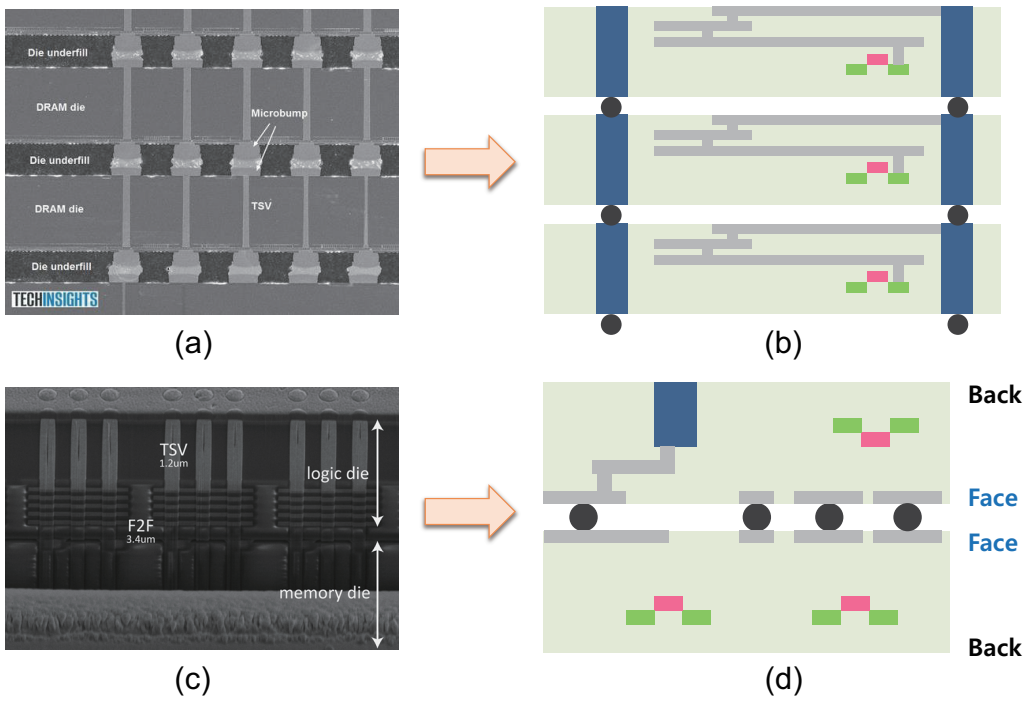


Figure 2: 3D ICs using TSVs and F2F bumps: (a) Actual 3D IC product using TSVs [76], (b) illustration of a 3-tier 3D IC, (c) 3D IC designed in F2F bonded style [32], and (d) illustration of a F2F bonded 3D IC with F2F bumps.

as TSV voids and cracks reduce the production yield (see Figure 3). In addition, manufacturing TSVs inside chips require the silicon substrate to be thinned (less than 100um). Handling thinned dies are challenging, and it becomes more challenging when technology scales and requires manufacturers to handle even thinner substrates.

In terms of the design side, TSVs are manufactured in a feature size that is significantly larger than regular transistors. In fact, typical TSVs are more than ten times bigger than standard cells. Thus, having more TSVs in designs means less silicon space for IP. In addition, TSV manufacturing induces significant stress to other components altering the performance of transistors. In silicon interposers, the unique interconnects that it provides cause signal integrity and power integrity problems to the ICs that are monted on it. Even in the system-level side, various challenges exist: First, I/O management issues arise. I/Os in 3D ICs must be aligned since I/Os on the top tier and bottom tier must be placed on the same coordinates. Second, multi-die logic partition and floorplanning become issues

due to the increased system complexity. Third, system-level reliability problems such as thermal or EMI issues also occur because ICs are now closer to each other. In addition to the challenges described above, many other 3D IC related challenges exist and must be conquered for reliable future electronics.

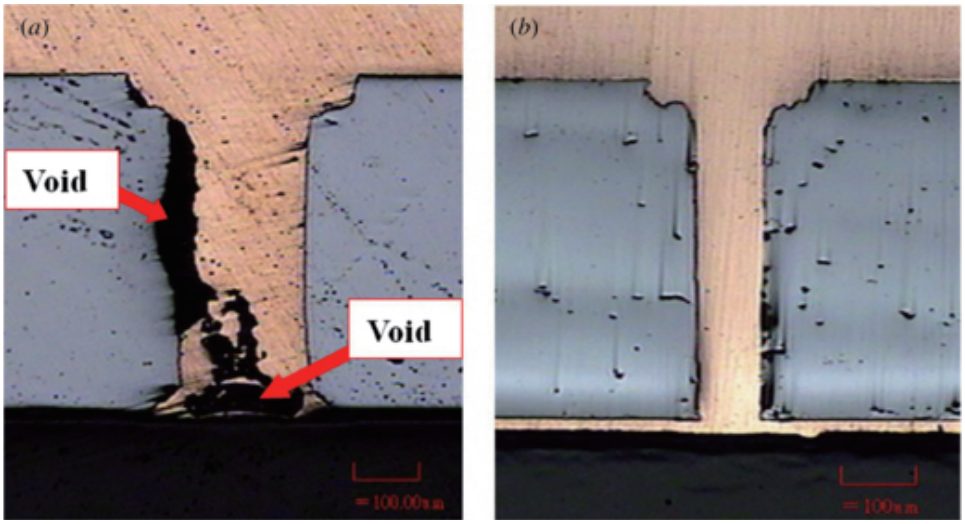


Figure 3: (a) TSV manufactured with voids, (b) TSV manufactured with no voids [13].

1.3 Scope of This Dissertation

This dissertation proposes co-design methodologies for reliable silicon-interposer-based 3D IC systems . In addition to this, it describes many other 3D IC related reliability issues for high-performance and low-power systems. Detailed contents include (1) system-level IR-drop analysis including 3D ICs and silicon interposers, (2) thermal analysis for 2.5-D based systems on silicon interposers, (3) full-chip level TSV-to-TSV coupling analysis in 3D ICs, (4) parasitics analysis and extraction for face-to-face bonded 3D ICs, and (5) 3-tier 3D IC design for more power reduction.

1.4 Organization and Contributions

This dissertation is organized in the following order to describe its contributions:

- In Chapter 2, two co-analysis studies are performed. First, a design methodology of co-analyzing IR-drop in 3D ICs, silicon interposer, and PCB is presented. IR-drop is an important issue in silicon interposers due to the thin metals that are used inside. Thus, it is important to study how significant IR-drop it will cause in systems. By proposing a holistic IR-drop analysis platform including 3D ICs and silicon interposers, the design turn-around-time and over design of PDN could be avoided. Second, a platform to co-analyze temperature in analog/digital mixed signal systems including silicon interposer is proposed. It was proven that integrated voltage regulators (IVR) could be embedded into ICs. However, their thermal characteristics have not been studied yet. Through our holistic platform, we analyze the thermal impact of IVRs and propose optimization methodologies to reduce temperature.

- In Chapter 3, it is shown how TSV-to-TSV coupling is different in ICs compared to packages and PCBs in both device level and full-chip level. In 3D ICs, the electrical characteristics of TSVs are different from that of TSVs in silicon interposers due to the I/O driver that drives the TSV. Therefore the coupling behavior becomes also different. Therefore, this chapter first studies the unique coupling mechanism of TSVs inside ICs and proposes methodologies to reduce coupling. Then, knowing from the unique coupling characteristics in TSVs, this chapter proposes an accurate methodology of performing multiple TSV-to-TSV coupling analysis and proposes optimization methodologies to reduce coupling in full-chip level.

- In Chapter 4, face-to-face bonded 3D ICs are studied and analyzed. When 3D ICs are bonded in F2F style, it introduces new parasitics due to the close distance between dies. Therefore, this chapter introduces what new parasitics exist in F2F bonded 3D IC structures. Then, it proposes a methodology of extracting these parasitics and study its impact in timing and power in full-chip level.

- In Chapter 5, the possibility of 3-tier 3D IC designs for more power reduction is

studied. Many previous studies showed how 3D ICs could lead to power reduction. This chapter shows how various 3D IC design techniques such as floorplanning, pin assignment, and block-folding contributes to more power reduction in 3-tier 3D ICs. In addition, 3-tier 3D ICs can be designed with various bonding styles. The impact of these various mixed bonding styles are also studied.

- In Chapter 6, the research in this dissertation is summarized.

CHAPTER II

CO-ANALYSIS METHODOLOGIES IN CHIP, PACKAGE, AND PCBS IN EMERGING TECHNOLOGIES

3D IC and silicon interposer technologies have emerged as two leading contenders for high speed, large-scale integration platform. 3D ICs using through-silicon vias (TSVs) have already been reported [36], and silicon interposer-based commercial product has also been released [16]. However, many design and analysis issues in silicon interposers and 3D ICs have not been delivered yet. For example, power delivery issues and thermal analysis methods still remain as questions for systems containing 3D ICs and silicon interposers. In addition, system-level analysis is more challenging than singular analysis because designers must handle multiple domain problems at the same time. Therefore, this chapter discusses issues and proposes methodologies to show how multiple-domain problems can be tackled for accurate analysis when systems are containing silicon interposers and 3D ICs.

The first part of this chapter is power delivery co-analysis. Silicon interposers use a very thin metal due to process issues. Comparing this with FR4 packages, it is less than 10% of the metal thickness used there. What makes it harder to design power distribution network (PDN) in silicon interposers is that it can use wide metal lines only that its width is limited to few tens of μm. It does not allow designing large metal planes for PDN while other packaging substrates support it easily. Thus, silicon interposers can cause a significant IR-drop noise in the PDN, and this can in fact affect power delivery to the 3D IC mounted on it. In order to accurately calculate the overall power delivery noise in the system level, it is necessary to simulate 3D IC, interposer, and PCB in a holistic fashion.

There have been several studies related to the co-analysis of chip-package and PCB.

However, there has not been any work that performs co-analysis of package, PCB, and a full transistor switching activity of 3D IC. [34] modeled PDN into small S-parameter blocks and connected them to obtain the whole PDN information of the system. However, it was only possible for periodic structures. [18] suggested to combine Laguerre Polynomials with the FDTD method to analyze the system PDN, but it had limits on simulating a very complicated PDN inside the ICs due to different aspect ratio between ICs and packages. [12] presented a co-simulation on DDR3 DRAM. However, power details inside the ICs were not provided. Therefore, the first part of this chapter discusses how severe the IR-drop noise is in silicon interposers. Then, the co-analysis methodology that calculates the IR-drop noise of the whole system with full transistor level power information details is presented. This research demonstrates the IR-drop results of a system, when silicon interposer is an alternative the organic packages.

The second part of this chapter is co-analysis for thermal impact. Low power is the essential keyword in modern system designs. For low power digital systems, dynamic voltage and frequency scaling (DVFS) is a well-known method to reduce power by adapting the voltage and frequency to changing workloads. To implement DVFS effectively, digital systems must be supported by voltage regulators that change power supply levels on nanoseconds.

Voltage regulators are used in many systems and are essential to provide power from energy sources to target systems. To implement a high-efficiency voltage regulator, inductor-based switching voltage regulators are commonly used. Conventional inductor-based switching regulators are operated at a relatively low switching frequency ($<$ 5MHz) and use bulky passive elements (e.g., SMT (surface mount) inductors and capacitors) for output filtering. Therefore, these voltage regulators are placed separately on the system board, limiting the systems to run in slow voltage adjusting capability [37].

An on-chip integrated voltage regulator (IVR) enables the effective implementation of DVFS. An on-chip IVR, operating at high frequency ($>$ 100MHz) does not require bulky

passive components (filter capacitor and inductor), allows the filter capacitor to be integrated entirely on the chip, places smaller inductors on the package (or on-chip), and enables fast voltage transitions at nanoseconds. Because of these advantages, several studies proposed various methodologies for IVRs [3, 21, 66, 80]. However, the primary obstacle faced in the development of IVRs is the integration of suitable power inductors. Recently, an early prototype of switched-inductor IVR using 2.5D chip stacking for inductor integration has been proposed [72] (see Figure 4).

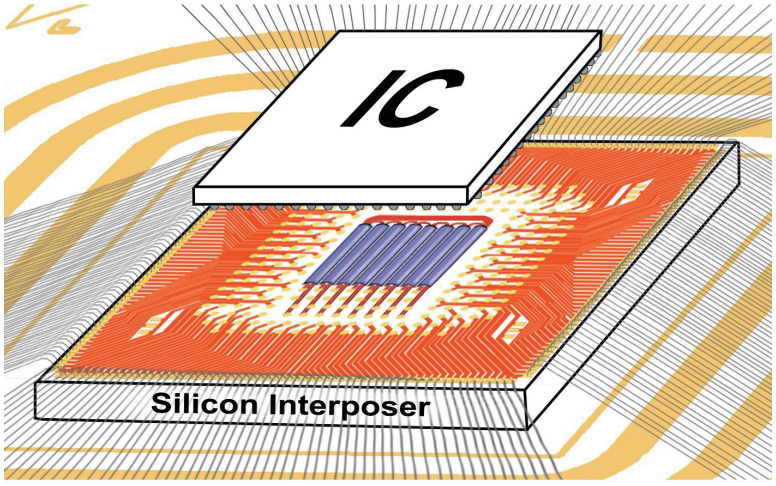


Figure 4: Diagram of a 2.5D integrated voltage regulator (IVR) chip stack. The IC consists of buck converter and load circuitry, and the silicon interposer contains the power inductor. The IC is flip-chip mounted on the silicon interposer using ball grid array, and wirebonds connect the silicon interposer and the IO.

Thanks to the recent development of these on-chip IVRs, the voltage regulators can be integrated inside the chip. However, when IVRs are integrated in the IC, they cause significant heat problems. The heat problems of voltage regulators in the system level were avoided when these regulators were placed separately on the system board. However, by placing these regulators inside the same IC, designers must consider the impact of a new heat source being added to the whole system. Currently, many tools exist to perform thermal analysis, but most of these tools focus on the analysis of package-level design. There exist several tools that can perform thermal analysis on the IC level design, but these tools do not describe how thermal analysis can be performed in an analog/digital mixed

system in GDSII layouts [58, 23, 82, 5].

Therefore, the second part of this chapter proposes a methodology of analyzing temperature of analog/digital mixed systems in GDSII-level details starting from the following sections. Using the proposed methodology, this research studies thermal impact on a 2.5D analog/digital mixed system with an IVR using silicon interposer [72] and demonstrate how critical the thermal problem is when the IVR is integrated in the IC.

2.1 A Co-Simulation Methodology for IR-drop Noise in Silicon Interposers

This section discusses the impact of IR-drop noise on silicon interposer. A system is designed that has an IC, an interposer and a PCB as in Figure 5 with the dimensions and details below. Due to the process issues, the width and thickness of the metal inside the interposer are limited. Here, the silicon interposer is assumed to have the metal thickness of 1μm, maximum width of 50μm, and minimum spacing of 50μm for PDN design. It is also assumed that the interposer has TSV in the height of 100μm and diameter of 20μm. The die size of the IC is 1mm $\times$ 1mm, silicon interposer 4mm $\times$ 4mm, and PCB 6mm $\times$ 6mm (metal thickness: 36μm). 81 power pins are distributed between IC and interposer in 100μm pitch, and these pins are connected with 30μm diameter C4 bumps. The system has 36 solder ball connection between the interposer and the PCB, and is distributed in 700μm pitch. Total power consumption is 1027mW, and 933.6mA flows through the system. One current sink was assigned at the middle of the IC model for worst case analysis.

Figure 6 shows the results. Ansys Siwave is used to simulate the system, and the results show that 17.08mV IR-drop noise occurs on the interposer and PCB, while an organic package (metal thickness: 18μm) and PCB shows less than 2.3mV of IR-drop. However, note that the maximum IR-drop generated by PCB is only 0.8mV. Thus, compared with the packages, silicon interposer causes significant IR-drop that must be managed properly.

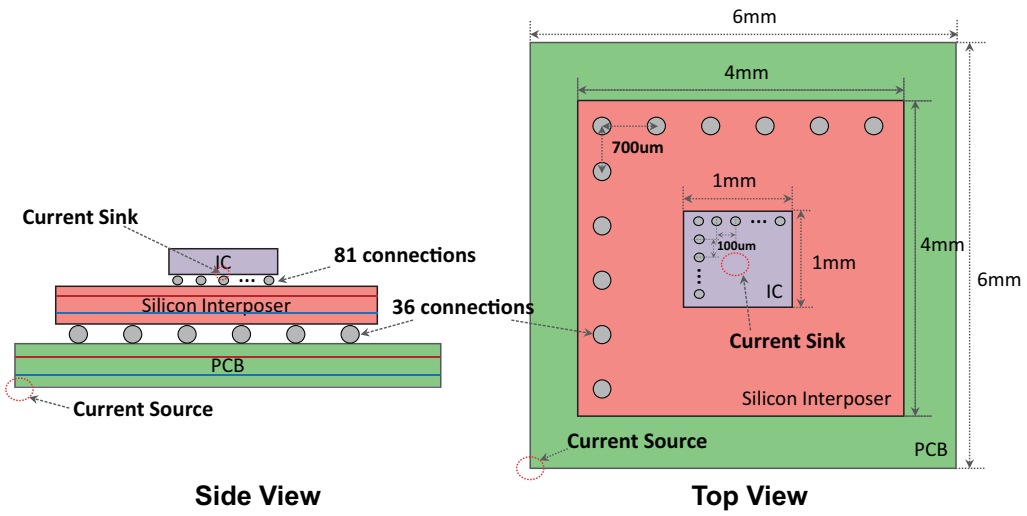


Figure 5: Side view and top view of the system simulated for IR-drop noise.

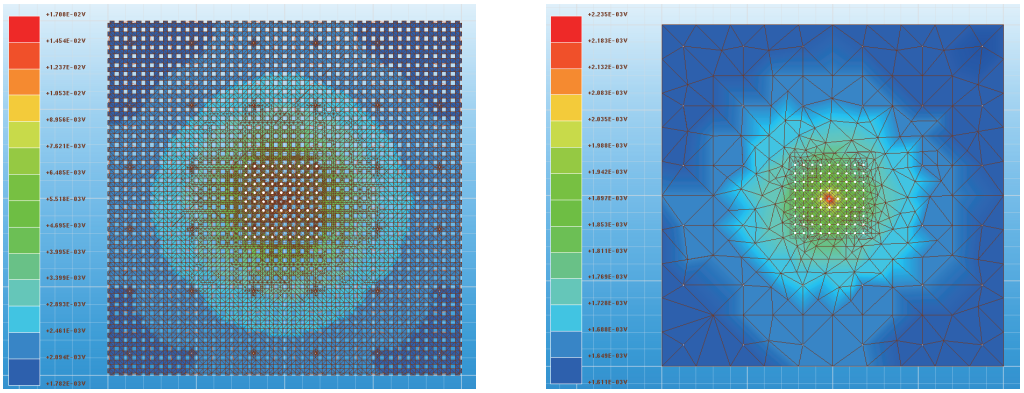


Figure 6: IR-drop Noise on (a): Si-interposer (17.08mV), (b): Organic package (2.24mV).

2.2 Interposer-3D IC Co-Simulation Methodology

This section describes the details of the proposed co-simulation methodology. Synopsys PrimeRail is the tool used for the co-simulation, and proper adjustments are made to implement the holistic platform. The design and modeling process diagram is shown in Figure 7, and the full details are described in the following subsections.

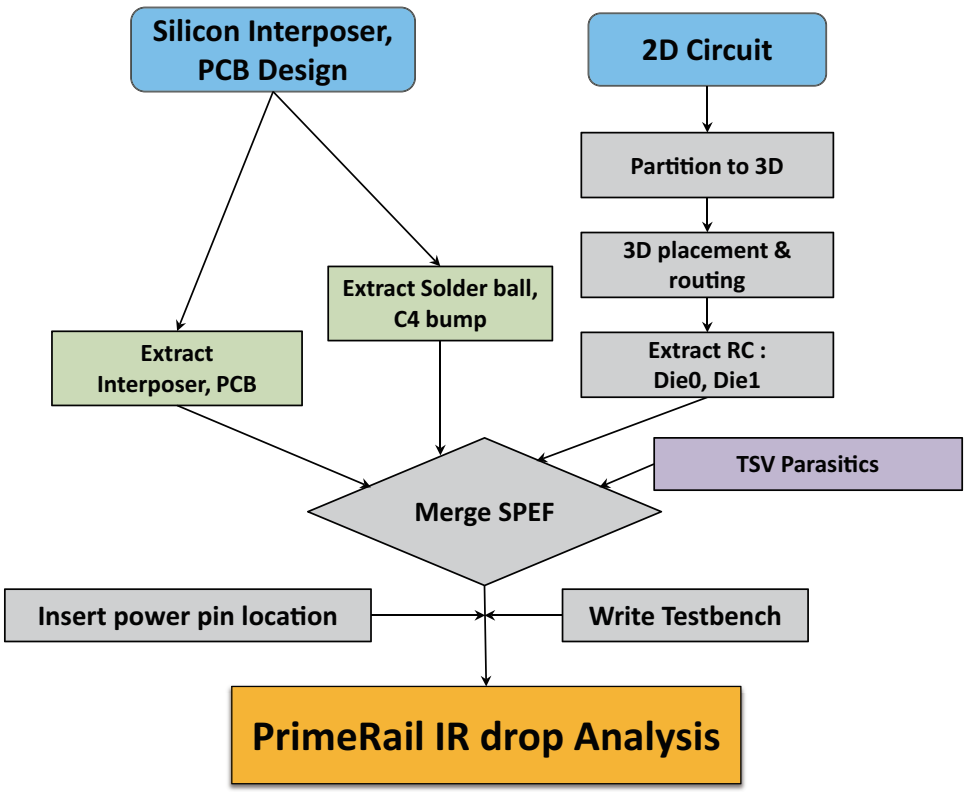


Figure 7: The proposed co-analysis design flow for IR-drop noise.

2.2.1 PDN Design of the 3D IC

To perform co-simulation of the whole system, PDN design of the IC is firstly needed. The design used in this chapter consists of a two-tier 3D IC that has face-to-back configuration as shown in Figure 8 (a). A peripheral PDN ring is designed using M1 and M2. M1 is used to supply power in standard cells, and M4 was used to support the vertical path. Details of the on-chip PDN are shown in Figure 8 (b). Nangate 45nm technology was used for this research. VDD is 1.1V, and TSVs in the 3D IC design has diameter of 5μm and height of 60μm.

2.2.2 PDN Design of the Interposer and PCB

For silicon interposer, PCB, and other interconnects, the design that has been made in Section 2.1 is reused (see Figure 5). To model the PDN of silicon interposer and PCB, a unit cell based SPICE method in [68] is used. Off-chip PDN design (interposer, PCB)

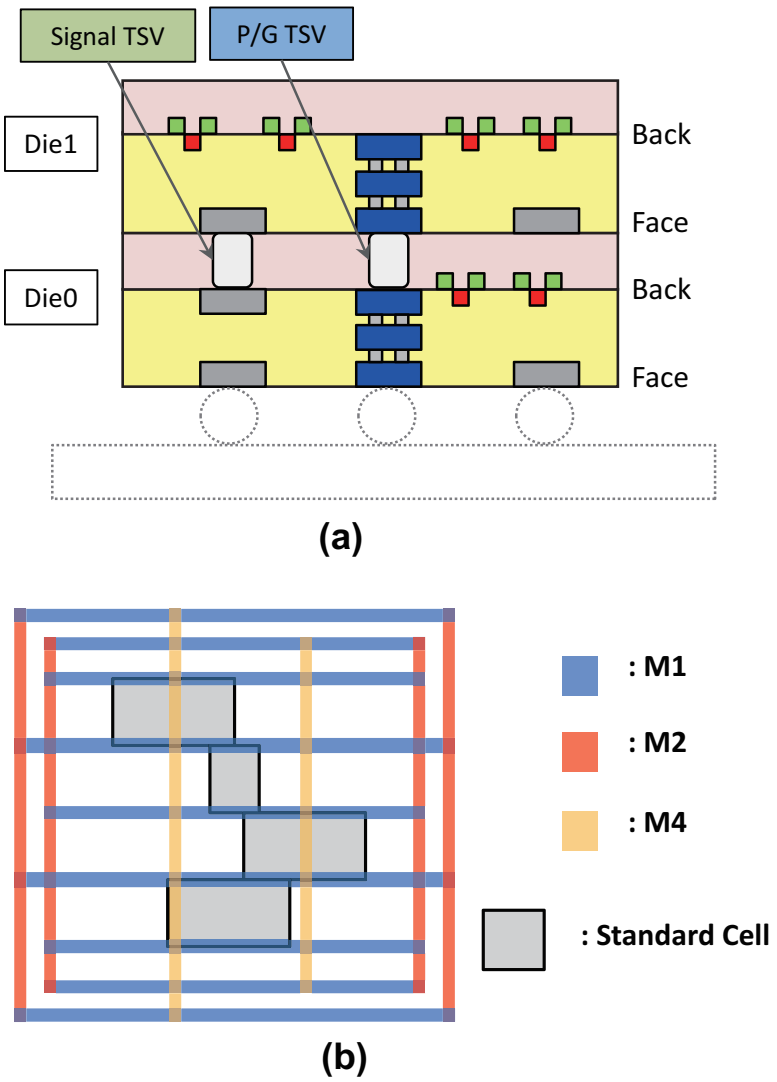


Figure 8: Details of the 3D IC PDN design (a): Stack information of the two tier 3D IC, (b): PDN design on the 3D IC.

could be split into array of unit cells as in Figure 9. Each unit cell describes a cluster of SPICE elements, and by connecting these together, the whole PDN can be reconstructed. Figure 9 shows a unit cell of 4×4 array, but other grid sizes are also possible, and this method can also be applied to irregular shaped PDNs. Each unit cell of silicon interposer, and PCB PDN represents a size of 100μm × 100μm. The resistance of each unit cell were extracted using Ansys Q3D Extractor. C4 bumps, TSV of interposer, and solder bump models were also made. SPICE values of these elements were also extracted using Ansys Q3D Extractor.

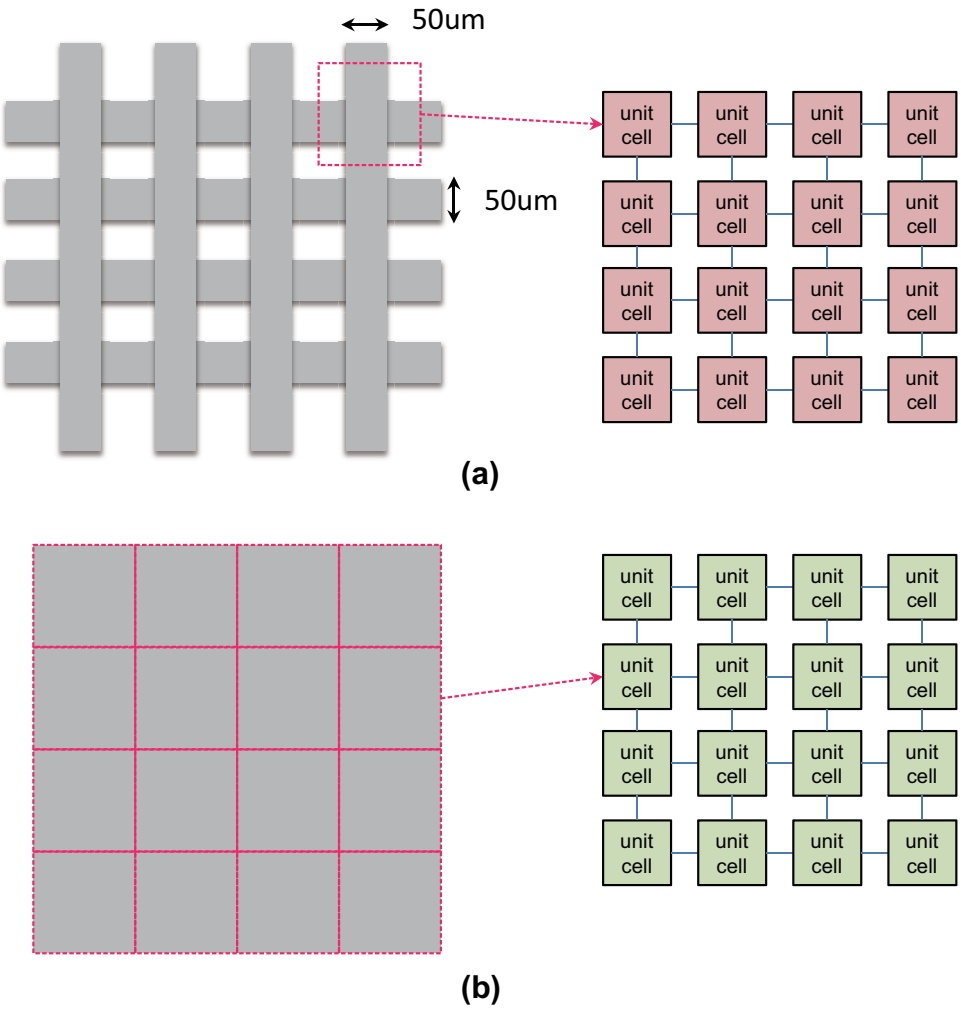


Figure 9: PDN modeling using unit cell model (a): Silicon interposer, (b): PCB.

2.2.3 Co-Simulation Methodology

Synopsys PrimeRail is a tool that is originally designed to analyze the PDN in ICs. It has a limitation of 15 metal layers that can be used. Therefore, if an IC design exists that uses less than 15 metal layers, additional layers can be added for extensions.

The proposed co-simulation methodology is shown in Figure 7. First, a 3D IC design is generated using 2D schematic. The 2D circuit is partitioned into several clusters, and each cluster represent each tier in 3D IC. The 3D IC design was performed using Cadence Encounter and in-house tools [33]. Then, standard cell placement and power/signal routing is performed. After routing and placement is done, the RC values of each tier are extracted

using Synopsys StarRC, and then merged into one SPEF (Standard Parasitic Exchange Format) file. In this file, all the P/G information (power rail, parasitic capacitance...etc) are gathered including geometry information of each metal layer inside the 3D IC.

Second, PDNs of silicon interposer and PCB are designed, and the information of each metal layer and interconnects are extracted. The extracted PDN information of the interposer and PCB are composed of SPICE elements and nodes connecting them. The extracted information is converted, then added into the same SPEF file that has the 3D IC information. To convert SPICE into the SPEF format, each SPICE elements are assigned a virtual width and length, and each node is assigned with a virtual location. In this study, the unit cell of a mesh PDN and a plane PDN both look like the same cross shape in SPEF file as Figure 10. Therefore, when these unit cell are combined together, the mesh PDN, and plane PDN would look like the same mesh shape in SPEF file. Using these converted information, the IC and the system components are logically connected in the SPEF file.

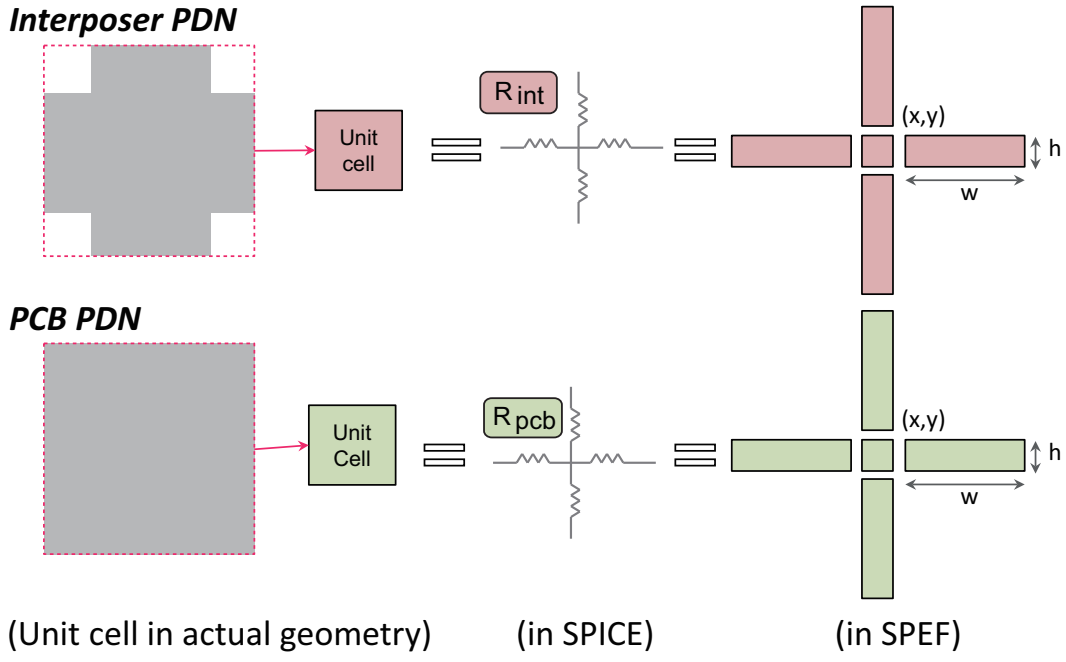


Figure 10: PDN unit cell translation from physical model to SPEF netlist (a): Silicon interposer, (b): PCB.

The silicon interposer and PCB are assigned to metal layers that has not been used for

routing in the IC design. Thus, it is important to leave a few metal layers empty during IC design. If the PDN of the IC consumes all 15 metal layers, then there would be no space left to insert the extracted system components in the SPEF file. Figure 11 shows the metal usage of this design. This research uses 6 metal layers for each tier of IC; one is for silicon interposer PDN, and one is for PCB PDN.

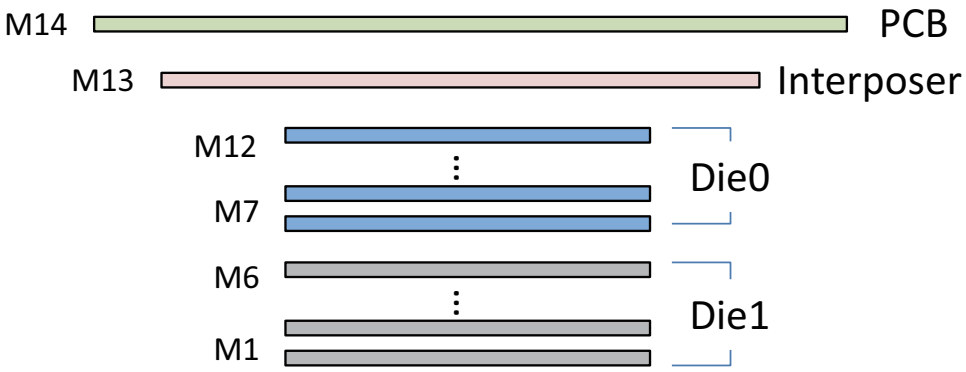


Figure 11: Metal layers used in Synopsys PrimeRail for IR-drop noise co-analysis.

Third, the SPEF file and other input files are inserted into Synopsys PrimeRail. Then, the simulation is performed. Two additional files are inserted into Synopsys PrimeRail: A LOC (location) file that has the layer number and the geometry information where the VDD source is located, and a verilog testbench that defines the vector activity of the standard cells.

2.3 Experimental Results

First, the unit cell method is validated to SiWave. Figure 12 (b) shows the IR-drop map of silicon interposer in SiWave when a current of 933.6mA is flowing, and the equivalent SPICE model in Figure 12 (a) using Keysight ADS. The maximum IR-drop between SiWave and SPICE is compared, and each voltages are 17.08 mV (SiWave), and 15.86 mV(ADS). The SPICE model shows good consistency with Ansys SiWave.

In Figure 13 (a), the result of a co-simulated PDN is shown. The 3D IC PDN is on the bottom, and the silicon interposer and PCB PDN mesh lays on the top as Figure 11.

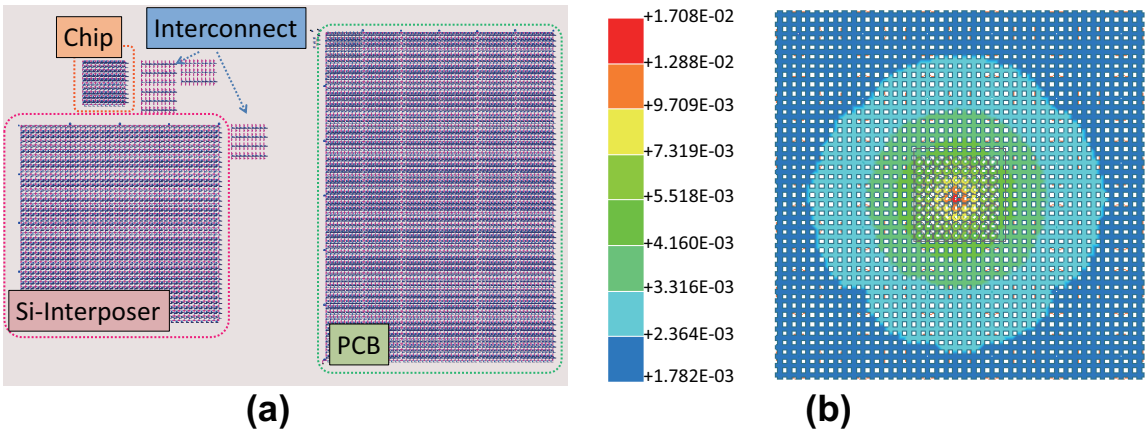


Figure 12: Validation of the unit cell model in comparison with Ansys SiWave (a): Keysight ADS (15.86mV, SPICE), (b): Ansys SiWave (17.08mV).

Figure 14 shows the top-down view of each layers. (a) shows the IR-drop map of the PCB, (b) shows the interposer, and (c), (d) show each tier. From Figure 14 (b), it is shown that silicon interposer generates a big IR-drop noise.

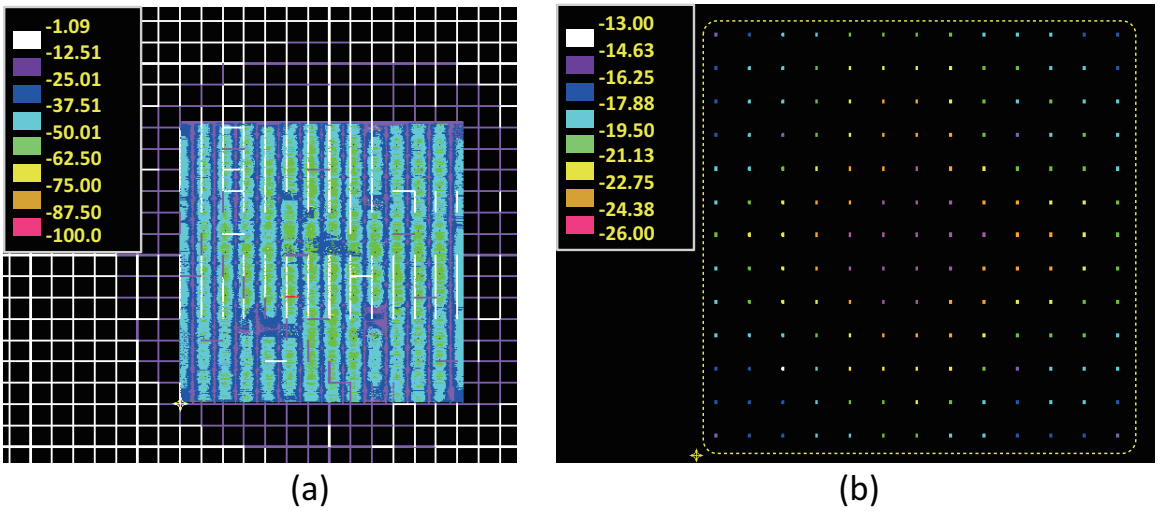


Figure 13: Co-simulated IR-drop result of FFT3 circuit in Synopsys PrimeRail (a): IC + Si-Interposer + PCB (full system), (b): C4 bumps

The importance of co-analysis is shown in Figure 13 (b), which describes an irregular IR-drop map of C4 bumps between interposer and IC. Without the gate level switching information, it is impossible to determine at which particular spot the IR-drop would be most severe, and which interconnect would supply how much current in which voltage.

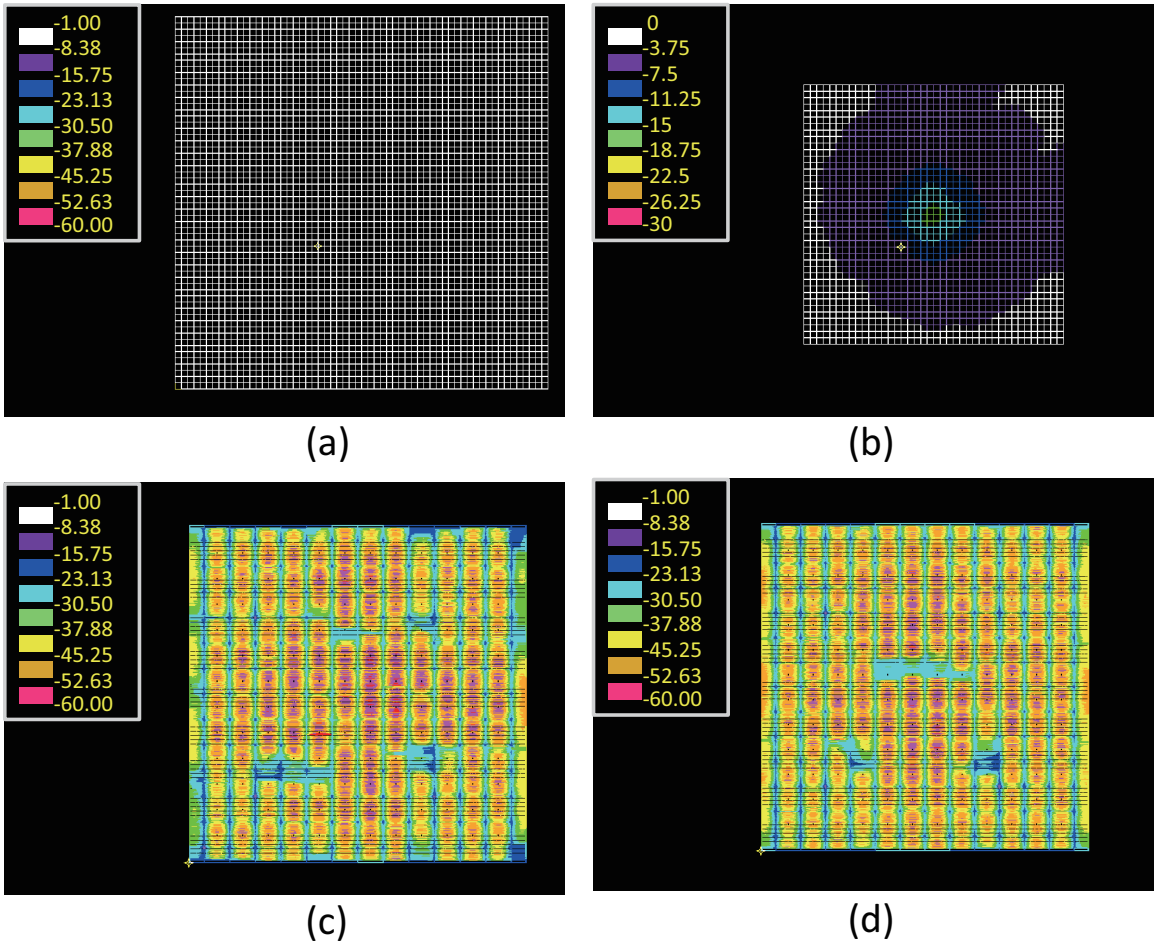


Figure 14: IR-drop map of each layers on the co-simulated PDN (a): PCB, (b): Si-interposer, (c): Die0, (d): Die1.

Figure 13 (b) is a valuable result, because this describes the actual detail on how much IR-drop is generated on each interconnect, which cannot be anticipated on separate analysis. Therefore, in IR-drop co-analysis, transistor level power details are very important.

To demonstrate the IR-drop co-simulation results of the system using silicon interposer, this study uses three FFT (Fast Fourier Transform) circuits that are described in Table 1. When separate analyses are done in FFT3 circuit, the IR-drop of the IC only PDN is 122.2mV, and IR-drop of interposer + PCB PDN is 35.0mV. However, when co-analysis is performed both on IC, interposer, and PCB simultaneously, the IR-drop is total of 147.7mV. The IR-drop of co-analysis is 9.5mV smaller than the separate analysis. 6.43% more IR-drop is overestimated in the separate analysis. The overestimation is also due to

the non-uniform switching activity of transistors in different locations, which can only be demonstrated in co-simulation. Table 2 details the results that have been performed with other circuits.

Table 1: Details of the circuits used in this paper

CKT	# of Gates	2D area	3D area	# Power TSV	# GND TSV
FFT1	140k	0.745mm^2	0.407mm^2	36	25
FFT2	297k	1.621mm^2	0.848mm^2	81	64
FFT3	616k	3.420mm^2	1.763mm^2	169	144

Table 2: IR-drop results comparison. PR stands for Synopsys PrimeRail

CKT	Power (mW)	IC (PR)	Int. + PCB (SiWave)	Co-anal.	Max. (ΣSep.)	Δ(ΣSep. - Co-analysis)
FFT1	558	94.9 mV	9.6 mV	103.8 mV	104.5 mV	0.7 mV
FFT2	1027	70.5 mV	17.1 mV	85.1 mV	87.6 mV	2.5 mV
FFT3	2137	122.2 mV	35.0 mV	147.7 mV	157.2 mV	9.5 mV

As the power consumption of the system increases, separate analysis overestimate more IR-drop than co-analysis [see Figure 15 (a)]. By this, it is expected to prevent more overestimated IR-drop by co-analysis when a system with a higher power consumption is analyzed. This is important because IR-drop is tightly connected to the total power consumption. Even with the same IR-drop, the total power loss of a system changes with the total power consumption. With an IR-drop overestimate trend like Figure 15 (a), the trend of overestimated power in higher power systems would be the square of Figure 15 (a), as in Figure 15 (b). Therefore, co-analysis is also necessary to estimate power correctly.

The ratio of IR-drop on silicon interposer to the total system is also high, compared to organic package. When using package between IC and PCB, IR-drop is less than a few mV, lower than 3% to the total IR-drop. However, when using silicon interposer, designers must consider a few tens of mV more. This is 16% to the total IR-drop, which is unnecessary in organic packages (see Figure 16).

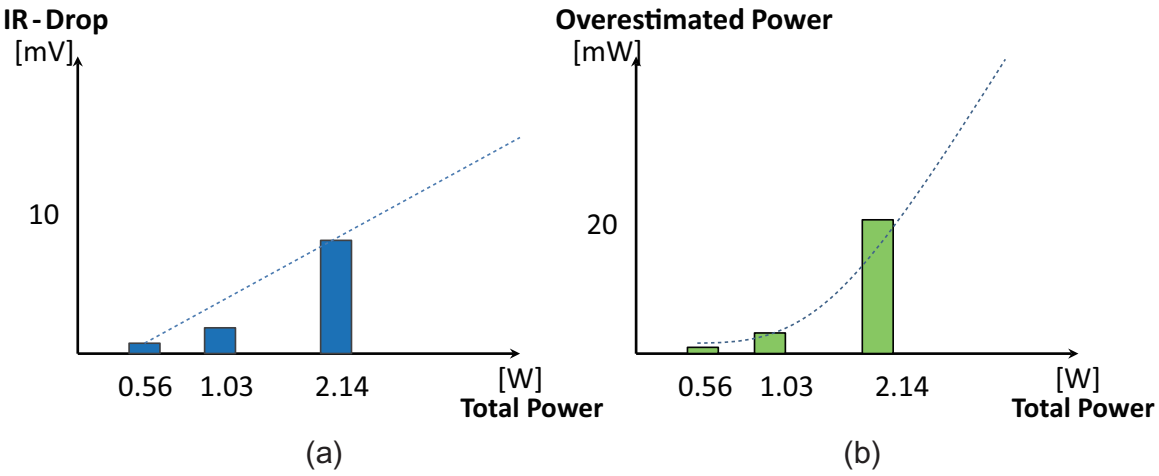


Figure 15: Benefits anticipation of co-simulation on (a): IR-drop, (b): Power saving.

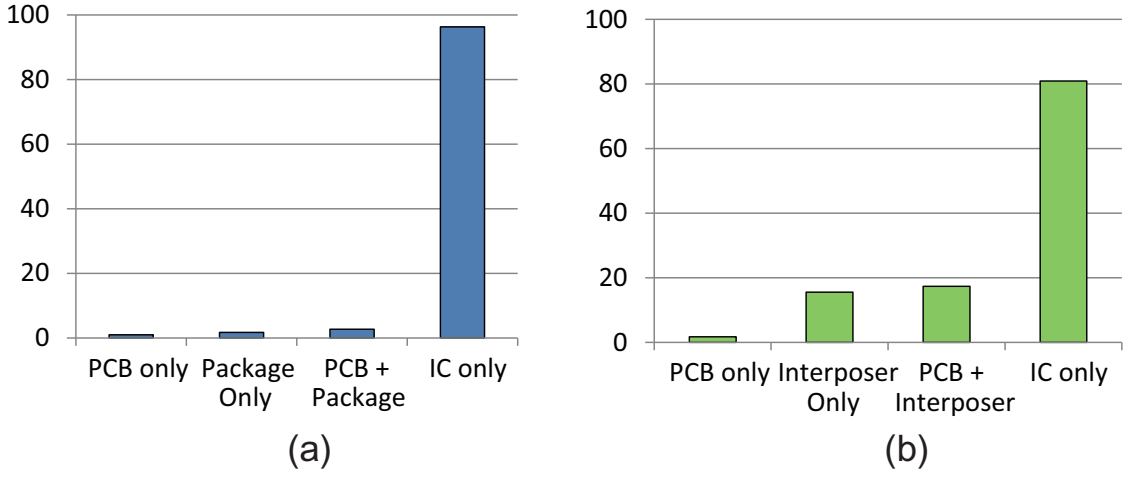


Figure 16: Ratio of each system components on IR-drop generation (Average of three circuits used on Table 2) (a): System with organic package, (b): System with Si-interposer.

2.4 Proposed Thermal Analysis Flow

This section proposes the design methodology for thermal analysis of analog/digital mixed designs. First, the full design methodology is described. Then, the detailed design methodology follows in the subsections. The main components of the design methodology are GDSII-level thermal analysis and power analysis.

2.4.1 GDSII Level Thermal Analysis

The following heat equation describes the steady-state temperature at a point $\mathbf{p} = (x, y, z)$ inside a 3D structure,

$$\nabla \cdot (k(\mathbf{p})\nabla T(\mathbf{p})) + S_{\mathrm{h}}(\mathbf{p}) = 0 \tag{1}$$

where k is thermal conductivity in $W/m \cdot K$, T is temperature in K, and S_{h} is volumetric heat source in W/m^3. By meshing the IC structure into elements as shown in Figure 17, the thermal model of Equation 1 is constructed for analysis. Each element, or thermal cell, represents a volume of specific length, width, and height. The height of a thermal cell is the same as that of each physical layer.

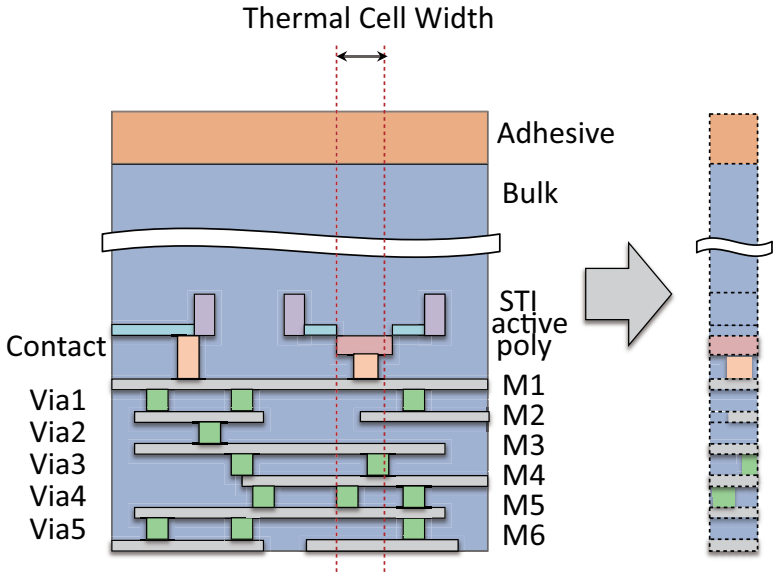


Figure 17: Example of thermal cells in a 6 metal layer IC. Total 17 layers of thermal cells are inside the dotted lines.

To solve Equation 1, boundary conditions on the six surfaces of the chip stack are required. Typically, a chip stack is very thin and flat and packaged inside molding materials. These molding materials are not good thermal conductors. Most of the heat flows from the bottom of the chip stack towards the heatsink. Thus, adiabatic boundary conditions are applied on the bottom and the four sides of the thermal structure. On the top side, a convective boundary condition is applied to model the heatsink.

The thermal analysis flow developed in this work is shown in Figure 18. Starting from the analog/digital mixed design netlist, the layouts are generated in GDSII format. A test-bench of the A/D mixed design is created from the netlist to perform power analysis of the functional blocks. In addition, the material density information of the layout is extracted from the layouts. The details of obtaining the information from the netlist (GDSII layout, power analysis, and material density) will be described in Section 2.4.2–2.4.4. Once the GDSII layouts, the power dissipation of each cells, and the material density information are obtained, the proposed design analyzer automatically generates the meshed thermal cells of the IC along with thermal conductivity and the volumetric heat source of each thermal cell.

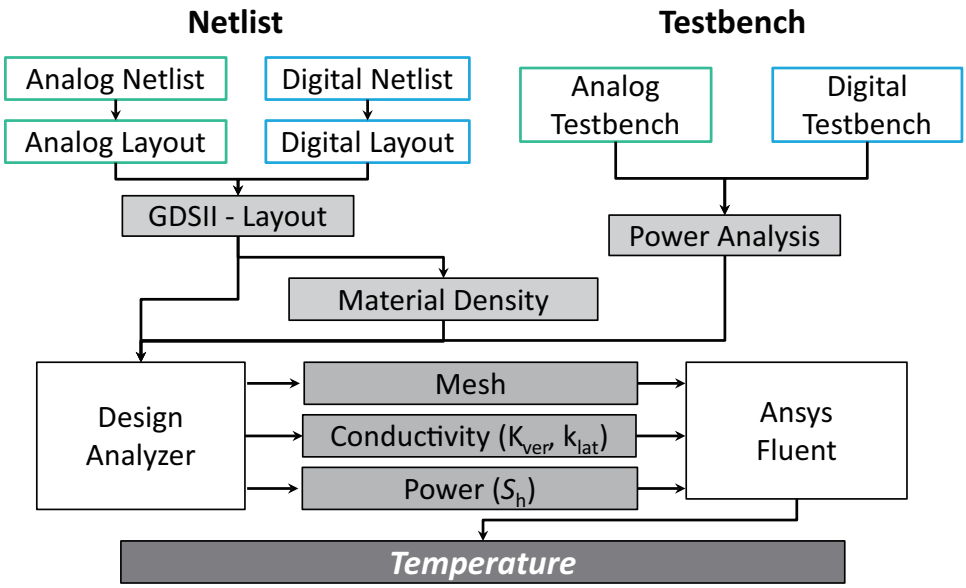


Figure 18: Proposed thermal analysis flow for the GDSII-level analog/digital mixed design.

A thermal cell may be composed of different materials. For example, in Figure 19, a thermal cell contains tungsten (for vias), copper, and dielectric. When a thermal cell is sufficiently small, an equivalent thermal conductivity based on thermal resistive model can be used [82]. Theoretically, if a thermal cell is very small, material inside the cell is homogeneous, and the thermal conductivity of the cell is isotropic. However, using a very small cell size requires high computing resources and a long runtime. Thus, for practical purposes, larger thermal cell sizes are used. Because of the typical structural geometries in GDSII layouts, the thermal conductivity of each thermal cell is anisotropic. The vertical thermal conductivity (k_{ver}) and the lateral thermal conductivity (k_{lat}) of a thermal cell consisting of N materials are computed by

$$k_{ver} = r_1 \cdot k_1 + r_2 \cdot k_2 + \cdots + r_N \cdot k_N \tag{2}$$

$$1/k_{lat} = r_1/k_1 + r_2/k_2 + \cdots + r_N/k_N \tag{3}$$

where r_i is the ratio of material i volume to thermal cell volume, and k_i is the thermal conductivity of material i. The proposed design analyzer computes r_i directly from the GDSII layouts of the chip stack.

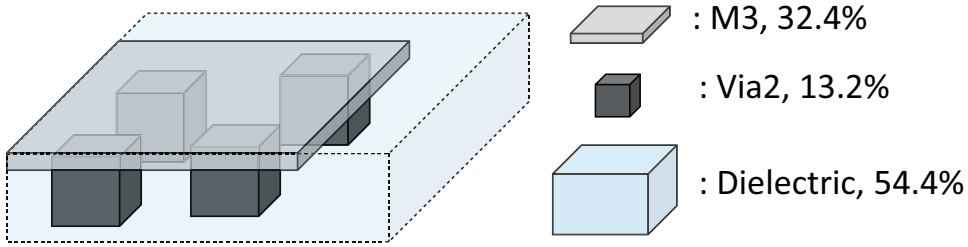


Figure 19: A thermal cell (dotted cube) with different material composition.

From the power dissipation and the location of each logic cell, total power dissipated in a thermal cell P_{cell} is calculated. Then, the volumetric heat source S_{h} is computed by

$$S_{\mathrm{h}} = \frac{P_{\mathrm{cell}}}{W_{\mathrm{cell}} \cdot H_{\mathrm{cell}} \cdot T_{\mathrm{cell}}} \tag{4}$$

where W_{cell}, H_{cell}, and T_{cell} are the width, height, and thickness of the thermal cell, respectively.

In this manner, Equation 1 is solved using Ansys FLUENT, a commercial thermal analysis tool. The meshed structure generated from the proposed layout analyzer is provided directly to FLUENT. In contrast, k_{ver}, k_{lat}, and S_{h} are fed into FLUENT through user defined functions because of position dependency. Finally, with the boundary conditions described earlier in this section, Ansys FLUENT is executed to obtain the steady state temperature of all positions in the chip stack. The proposed design flow can also handle multi-chip stack 3D ICs and chip stacks on silicon interposers.

2.4.2 Analog/Digital Mixed Thermal Analysis - Layout

The proposed thermal analysis flow requires GDSII-level layouts. From the analog/digital mixed netlist, the netlist is separated into analog and digital parts. Then, the analog layout is drawn using Cadence Virtuoso, and the digital layout is generated using Cadence Encounter. Finally, these two layouts are merged into one GDSII file.

2.4.3 Analog/Digital Mixed Thermal Analysis - Power Analysis

The proposed power analysis flow is shown in Fig 20. The power analysis is an essential step in the proposed thermal analysis flow, because the power of each transistor is the heat source S_{h} in the thermal analysis. The power analysis is separately performed for the digital and the analog parts. For the digital part, once the layout is generated in Cadence Encounter and saved in DEF or GDSII format, the parasitic resistance and capacitance of nets are extracted in SPEF format. In addition, Mentor Graphics Modelsim is executed for the testbench of the digital netlist to generate the switching activity of each logic cell in VCD format. Then, Synopsys PrimeTime PX is used to perform static power analysis and report power dissipations of logic cells. By stitching the power dissipation and the location of each cell using the DEF file, the power information of all locations in the layout is obtained.

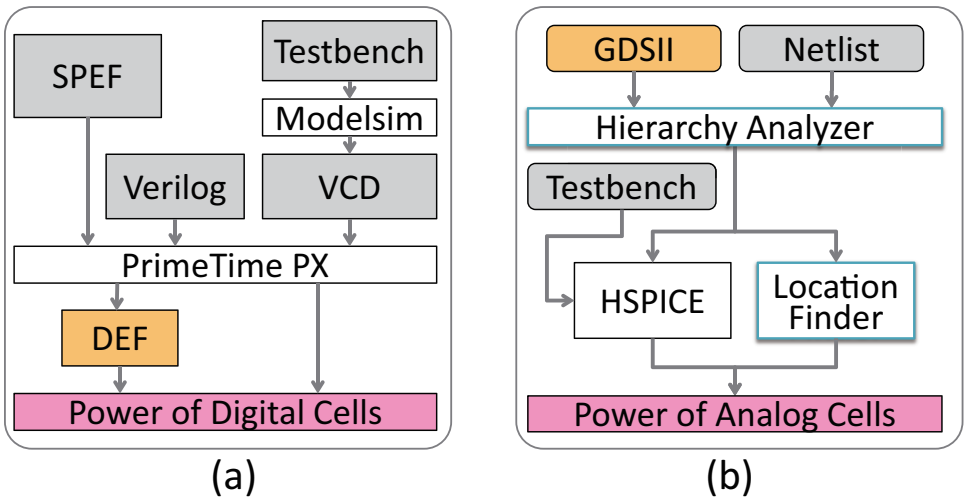


Figure 20: Power analysis flow of (a) digital design, and (b) analog design.

For the analog part, the design cells for power analysis are first chosen using the proposed Hierarchy Analyzer. An analog netlist may have multiple design hierarchies, from the high-level function blocks to transistor-level blocks. Therefore, it is important to decide which level of hierarchy is analyzed. Algorithm 1 describes the proposed algorithm. From a given netlist and the corresponding layout, a hierarchy tree of the netlist and the layout is constructed. Then, starting from the root cell (the highest hierarchy) of the netlist, the hierarchy between the netlist and the layout is compared. If the cell name in the corresponding layout hierarchy tree matches the cell name in the netlist, it descends down one cell and proceed with the same process. If the netlist and layout name matches to the lowest hierarchy, the lowest hierarchy cell is chosen for power analysis. If there exists a cell in the netlist with unmatched hierarchy in the layout, the parent cell for power analysis is selected. Figure 21 shows an example of how the proposed Hierarchy Analyzer works. Once the Hierarchy Analyzer chooses which design cells to perform power analysis, HSPICE is used to run power simulation with the testbench and the proposed Location Finder to search the location of each design cell in the layout.

Algorithm 1: Hierarchy Analyzer

Input : Netlist, GDSII layout
Output: List of chosen cells for power analysis

```
Construct a hierarchy tree of the netlist;
Construct a hierarchy tree of the layout;
Start from the root cell in the netlist hierarchy tree;
while Netlist hierarchy tree do
    Compare the netlist hierarchy tree cells from layout hierarchy tree cells;
    if A cell name in netlist hierarchy tree matches layout hierarchy tree cell
    name then
        if Last of hierarchy then
            Stop descending, choose the cell, and move to next branch;
        else
            Descend to it's child cell;
        end
    else
        Select the parent of the current cell and move to next branch;
    end
end
```

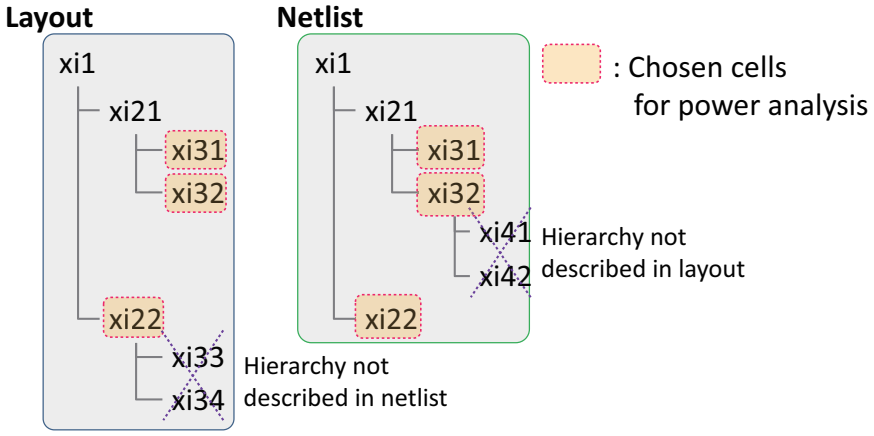


Figure 21: An example of Hierarchy Analyzer on a netlist, choosing analog cells for power analysis.

2.4.4 Analog/Digital Mixed Thermal Analysis - Material Density Library

In Section 2.4.1, it was explained that the proposed design analyzer provides k_{ver} and k_{lat} of each thermal cells to Ansys FLUENT using the material density information of the layout. However, both analog and digital parts consist of multiple design cells that are repeatedly used in the layout (e.g., standard logic cells). To reduce the computation time of these

repeating cells, a look-up table of material information is built so that the proposed design analyzer does not analyze same analog/digital design cells repeatedly. The look-up table is used to compute the material density and thermal conductivity covered by the area of the cell. Whenever the design analyzer encounters a cell that is described in the look-up table, it refers to the information in the look-up table.

2.5 2.5D Integrated Voltage Regulator using Magnetic-Core Inductors on Silicon Interposer

This section describes the integrated voltage regulator that will be analyzed using the proposed analysis flow. The integrated voltage regulator consists of the silicon interposer and the IVR chip, which contains the buck converter, control circuitry, and a network-on-chip (NoC) Load. The power inductor for the IVR is integrated on the silicon interposer.

2.5.1 Basic Structure of the Integrated Voltage Regulator

Figure 4 shows the complete 2.5D chip stack of the integrated voltage regulator. An IC, fabricated in IBM's 45nm SOI process, contains buck converter circuitry, decoupling capacitance, and a realistic digital load. This IC is flip-chip mounted onto an interposer that holds custom fabricated coupled power inductors for the buck converter while breaking out signals and the 1.8V input power supply to wirebond pads on the perimeter of the interposer.

The control circuitry occupies 0.178mm^2, while the bridge FETs occupy 0.1mm^2. The controller is designed to accommodate any number of inductor phases up to eight, and to provide a fast non-linear response to transients, allowing a reduction in the required decoupling capacitance on the output voltage [73]. Also, residing on the IC is a 64-tile NoC consisting of four parallel, heterogeneous, physical network planes with independent frequency domains. The NoC provides realistic load behavior and supports experimentation on supply noise and DVFS. A total of 48nF of deep-trench (DT) and thick oxide MOS capacitance decouples V_{OUT} and occupies 0.40mm^2, while 21nF of DT occupying 0.52mm^2

decouples the 1.8V input supply to compensate for the large PDN impedance.

2.5.2 Power Inductor inside the Integrated Voltage Regulator

A part of eight coupled power inductors shown in Figure 22 are fabricated on the silicon interposer such that one terminal of each inductor connects to a pair of V_{BRIDGE} C4 receiving pads, while the opposite terminals are shorted and connected to several pads across the interposer for distribution of V_{OUT}. The inductor topology is an elongated spiral with a Ni-Fe magnetic core encasing the copper windings on the long axis [38] [78]. The inductor fabrication involves successive electroplating deposition of the bottom magnetic core, copper windings, and top magnetic core. A hard-baked resist layer provides physical support to the top magnetic core and has a gentle taper to the sidewalls so that the top core arches over the windings without any abrupt transitions that would cause undesirable micromagnetic effects. The inductance decreases and resistance increases with frequency due to eddy currents, skin effect and domain wall motion.

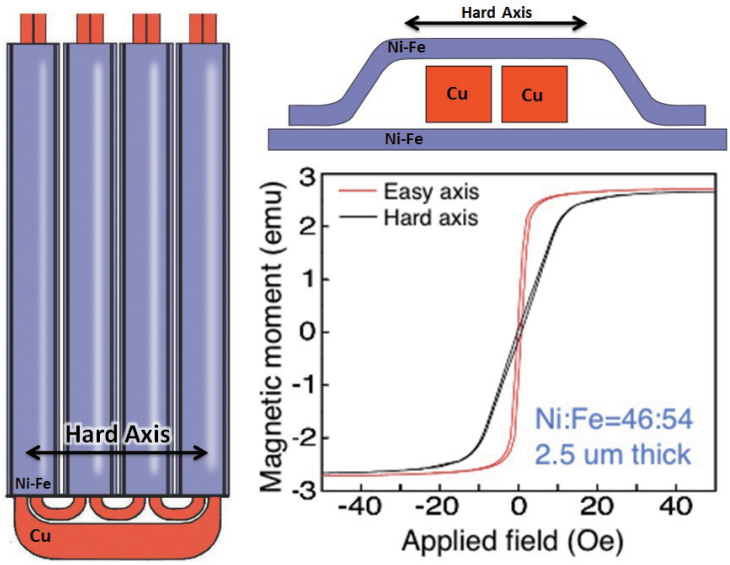


Figure 22: Top view of a part of eight single-turn, coupled power inductors (left), cross-section of magnetic cores and windings (top right) and magnetization curves for the Ni-Fe core material (bottom right).

2.5.3 Efficiency of the Integrated Voltage Regulator

Efficiency versus load current for the IVR is shown in Figure 23. Efficiency peaks at 74% with input voltage of 1.8V, conversion ratio of 0.61, switching frequency of 75MHz and load current of 3A. The FEOL current density is 10.8A/mm^2, which is defined as load current density divided by the FEOL area of the switches and controller, likewise the silicon interposer current density is 0.94A/mm^2, which is defined as load current divided by the total inductor area, 3.2mm^2. At peak efficiency, inductor DC and AC losses contribute approximately 26% and 48% of the total power loss, respectively, while switching and conduction of the bridge FETs contribute 25%. The peak current density occurs at 5.4A and efficiency of 66%.

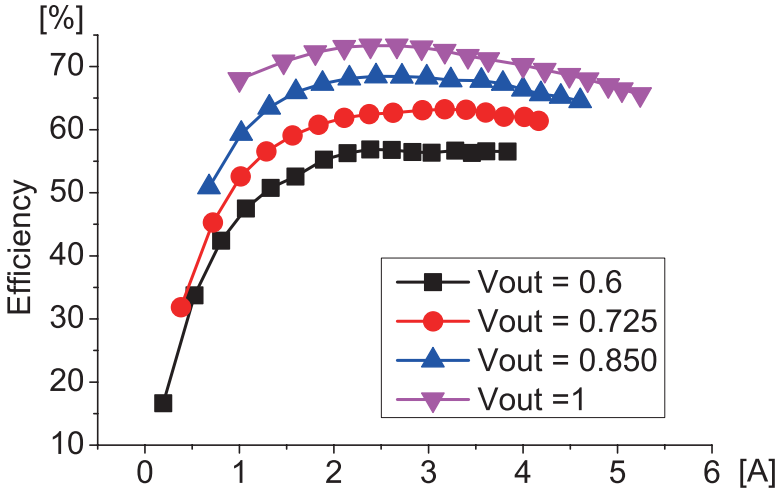


Figure 23: IVR efficiency as a function of load current at 75MHz switching frequency.

2.6 Thermal Analysis of the 2.5D Integrated Voltage Regulator

This section performs thermal analysis to the IVR. The physical dimensions are described first, then the analysis is followed.

2.6.1 Dimensions and Power Consumption of the Integrated Voltage Regulator

Figure 24 shows the structure analyzed in this study. The size of silicon interposer in the IVR is 6mm$\times$6mm, and the thickness is 720μm. An 8 cross-coupled inductor is designed

on the silicon interposer. The inductor footprint is 1.2mm$\times$1.2mm, and the metal used for the inductor is 5μm thick and 40μm wide. The size of the chip is 4mm by 4mm, and the thickness is 380μm. NoC is placed on the middle of the chip, and the buck converter is placed next to the NoC. I/Os, decoupling capacitors and peripheries are placed on the boundaries of the chip. 75μm C4 bumps are used to connect the silicon interposer and the chip, and epoxy underfill fills the empty space between the chip and the silicon interposer. A 3mm copper heat sink is assumed to be placed on the top of the chip. The power inductor is placed beneath the NoC due to routing issues.

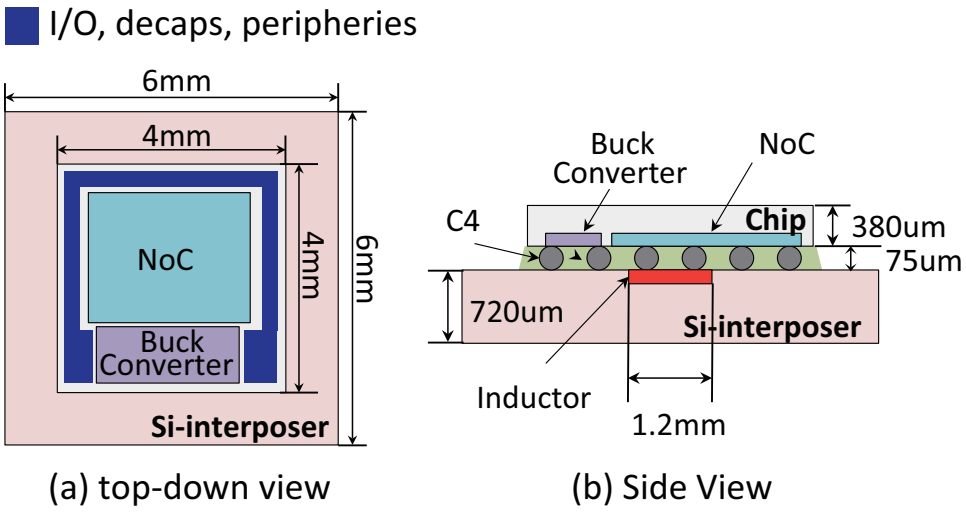


Figure 24: (a) Top-down view, (b) side view of the IVR.

From the input voltage V_{in}=1.8V, IVR can be operated in many different output voltage and load conditions shown in Figure 23. Therefore, this study assumes the IVR is operating in V_{IN}=1.8V, and V_{OUT}=1.0V. The temperature of each blocks when the load current changes from 1A to 5A are reported. Therefore, the power range in this study is from 1W to 5W. This study mainly focuses on buck converter, NoC, PDN of the chip, and the inductor because these are the most important blocks in thermal analysis. Table 3 reports some power numbers consumed by these important blocks for reference.

Table 3: Power consumption numbers of some blocks from the measurement.

	Low Current (1W)	Peak Efficiency (2W)	High Current (5W)
Buck Converter	83.55 mW	111.3 mW	305.5 mW
Inductor	323.7 mW	425 mW	1133 mW
NoC	1000 mW	2000 mW	5000 mW
PDN	45 mW	180 mW	1125 mW

2.6.2 Thermal Analysis of Essential Design Blocks

The thermal analysis starts by analyzing the temperature of design blocks assuming each blocks are operated separately. Figure 25 shows the thermal map of each design blocks. For the NoC, the maximum temperature rises up to 70.82°C when consuming high power (5W). Each NoC tiles show similar temperature map because NoC connects 64 symmetric digital blocks. For the inductor, the maximum temperature rises up to 77.3°C. Due to the fact that the inductor consumes high power in a relatively small footprint, the inductor is the hottest block of this 2.5D system. For the buck converter, the maximum temperature is 54.49°C. The IVR consists of one controller, and eight power drivers that are connected to the eight inductors. The hot spots in the buck converter is the eight power drivers. Decoupling capacitors and other circuitries exist between the power drivers. Therefore, a temperature valley is created between the hot spots of the power drivers. These temperature valleys reduces the temperature of the buck converter.

From 1.0W to 5.0W, temperature of each blocks were measured assuming each blocks are operating separately. Figure 26 shows the graph of temperature increase on each blocks. When consuming (generating) 5W, the highest temperature rise occurs from the inductor. Notice that there is a factor that can contribute to the temperature rise in the IVR. Power Distribution Network (PDN) of the IVR chip is a path where the input power (V_{IN}, I_{IN}) must flow before reaching to the buck converter. Since the buck converter and the NoC is integrated together, the PDN is above both the NoC and buck converter. However, the size of PDN is same as the total chip size (4mm$\times$4mm). Due to this, the temperature rise by

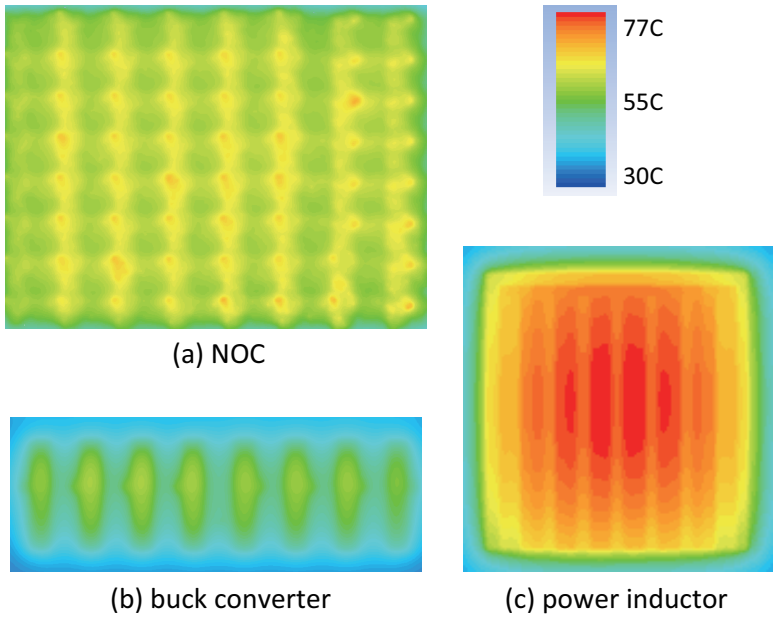


Figure 25: Thermal maps. (a) NoC, (b) power inductor, (c) buck converter when generating (= consuming) 5W.

the PDN is not so severe.

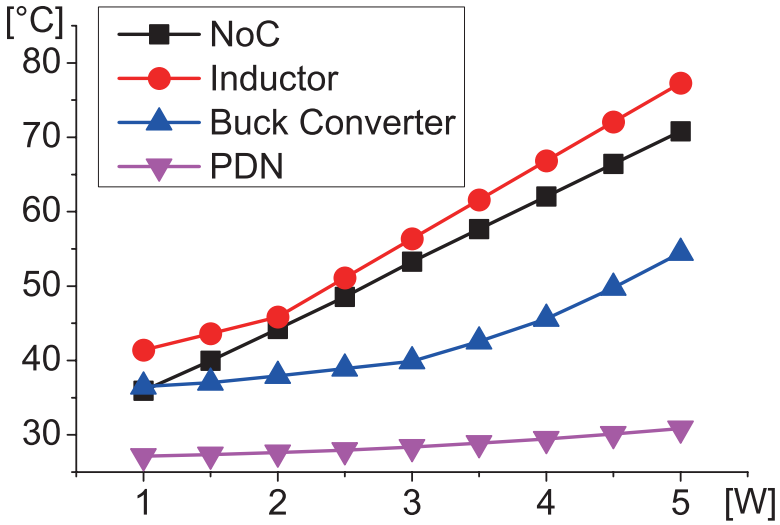


Figure 26: Temperature of each blocks in the IVR.

2.6.3 Factors Affecting Temperature Rise on Each Design Block

This section investigates what are the factors that affect to the temperature rise in NoC, and the buck converter.

2.6.3.1 *Factors Affecting Temperature Rise in the NoC*

Like in Figure 24 (a), NoC is surrounded by several other blocks. Buck converter is located next to the NoC, PDN is located above, and the inductor is placed beneath the NoC. Therefore, based on some scenarios (see Figure 28), this study investigates what are the most critical factor that affects temperature increase of NoC. These scenarios are:

1. NoC only. No surrounding block generates heat.
2. NoC and PDN. Buck converter designed on the same chip, but assumed to be far away from NoC.
3. NoC and buck converter designed on the same chip, sharing the same PDN.
4. NoC and inductor. Inductor placed below the NoC.
5. All components placed together (Figure 24).

From Figure 27, it is seen that the inductor beneath the NoC impacts to a high temperature, but buck converter designed with NoC at the same chip [see Figure 28 (c)] hardly affects to a temperature rise. The PDN impact a small temperature rise to the NoC. Figure 29 shows a temperature map when all components are placed together. A big thermal coupling occurs between the inductor and the NoC. Therefore, thermal coupling between the system and the inductor is a critical factor in IVR.

2.6.3.2 *Factors Affecting Temperature Rise in the Buck Converter*

Knowing that buck converter and NoC has a minor thermal coupling effect to each other, this study compares the following scenarios to analyze which scenario affect temperature increase the most in the buck converter. These are:

1. Buck converter only. No surrounding block generates heat.
2. Buck converter and the PDN.

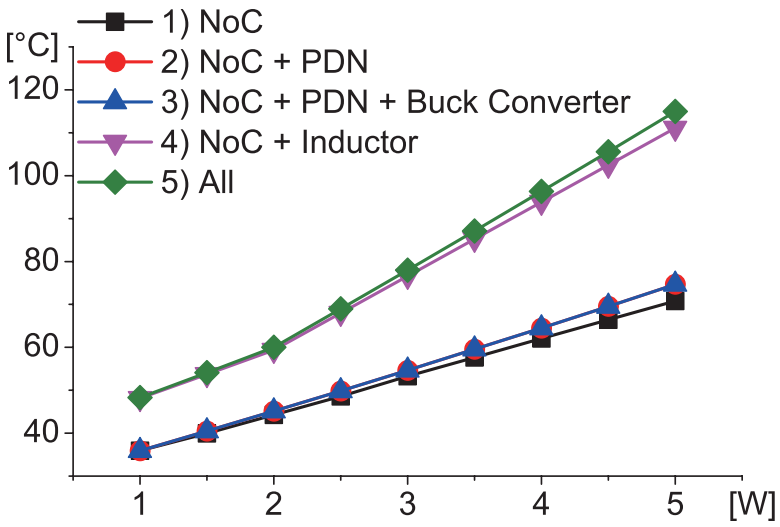


Figure 27: Temperature of NoC on different analysis scenarios.

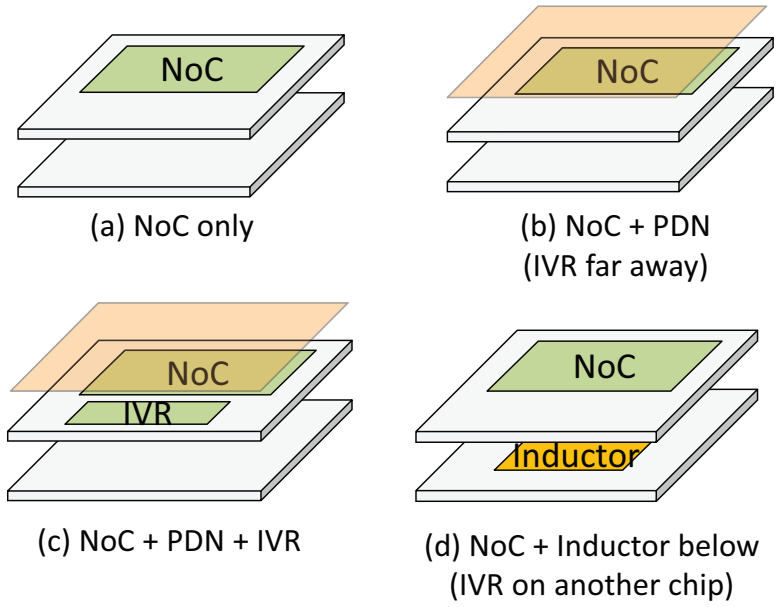


Figure 28: 4 Scenarios for NoC temperature analysis.

3. All components placed together (The real chip).

Figure 30 shows that only the PDN which is located above the buck converter affects temperature increase in the buck converter. Inductor has minor effect due to the far distance between the inductor and the buck converter.

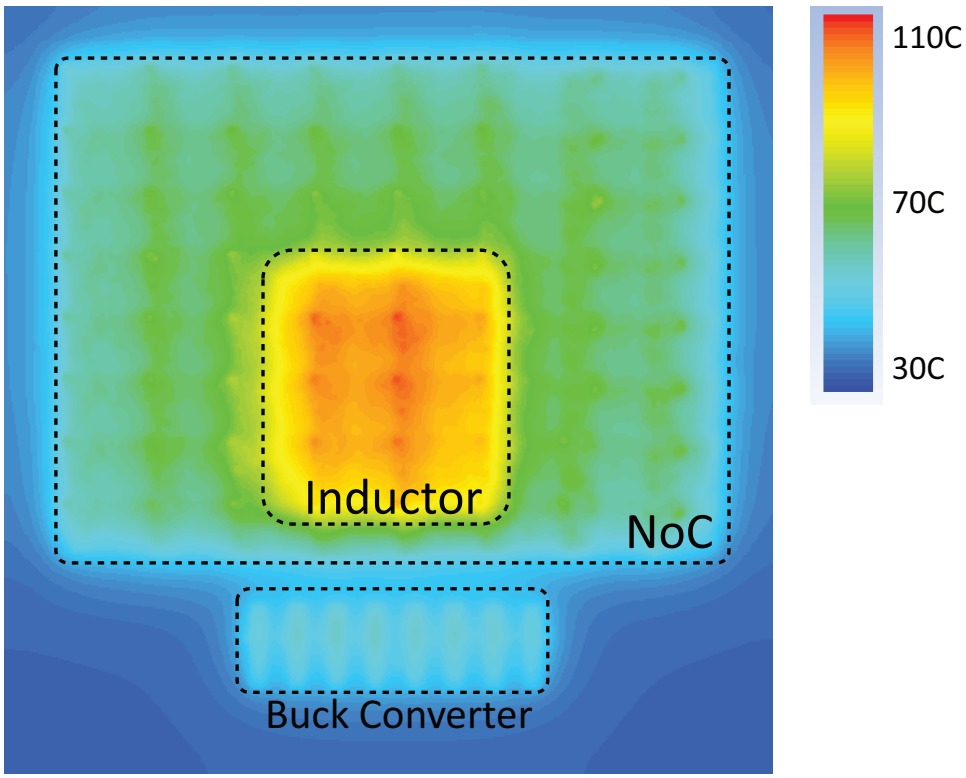


Figure 29: Thermal map of the IVR full chip when operating at 5W.

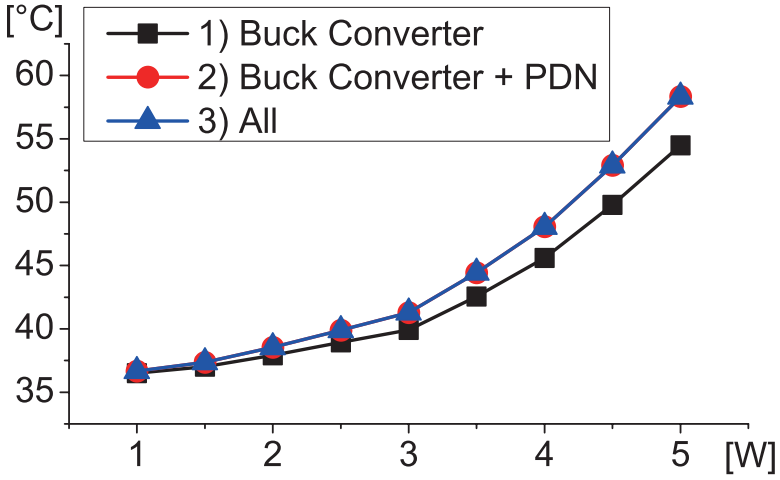


Figure 30: Temperature of buck converter on different analysis scenarios.

2.6.4 Thermal Coupling Between NoC and the Buck Converter

From section 2.6.3.1, it was shown that the thermal coupling between the NoC and the buck converter is minor. Here, this study further analyzes about some other impacts. A test was developed as shown in Figure 31. Starting from where the NoC and buck converter are

close to each other but do not overlap, the distance between these two blocks was increased and the highest temperature was measured. Here, three different scenarios were analyzed. With the same 5W generating buck converter, three different (5W, 3W, and 1W) loads were inserted to NoC for different temperatures. For reference, the maximum temperature of the 5W buck converter is 54.49°C, and the maximum temperature of 5W, 3W, and 1W NoC is 70.82°C, 53.27°C, and 35.94°C respectively. From Figure 31, it is seen that the maximum temperature is hardly affected by the distance between two blocks.

The impact of distance between the NoC and buck converter is almost negligible. This is because of the following reasons. First, the IVR is not a dense power consumer. Between the 8 power drivers, there exist low power consuming components that create a temperature valley between the high power drivers (section 2.6.2). The heat valley in the buck converter reduces thermal coupling between the buck converter and the NoC. Second, The hotspot of the 2D NoC is not on the periphery, which meets the IVR directly. Lastly, the heatsink is attached on the top of the IVR chip. Therefore, majority of the heat flows to the vertical direction than the lateral direction. Therefore, changing the distance between NoC and buck converter do not have a big impact on temperature increase. Figure 32 shows the temperature map when the distance between NoC and the buck converter is 0um and 100um.

2.6.5 Thermal Coupling Between NoC and the Inductor

Inductor beneath the NoC has a huge impact on temperature rise inside the IVR. This section further analyzes this impact by changing the overlapping distance between the NoC and the inductor. Like in Figure 33, three different scenarios are simulated and the highest temperature was measured. These scenarios are

- 5W consuming NoC + 5W generating inductor
- 3W consuming NoC + 3W generating inductor
- 1W consuming NoC + 1W generating inductor

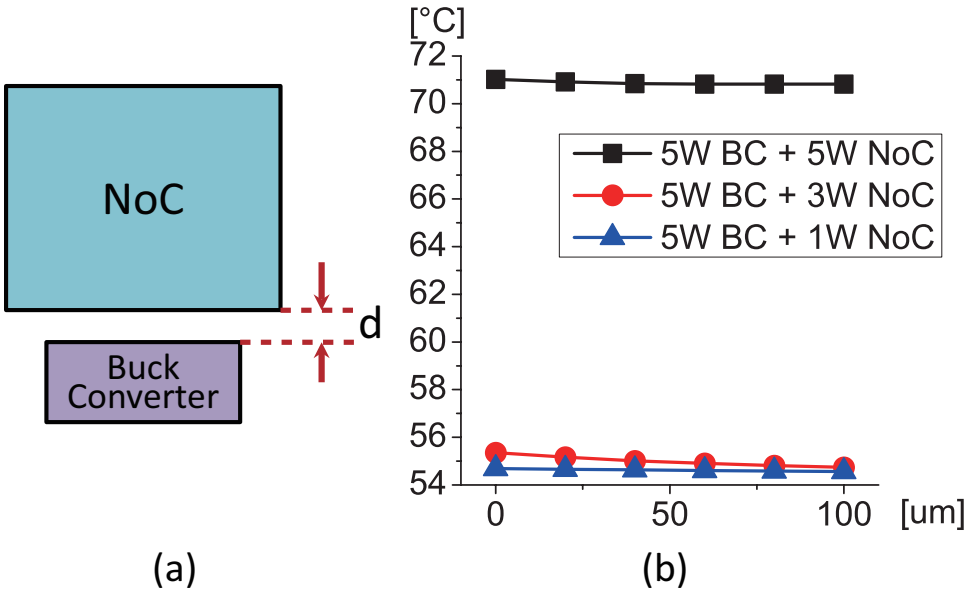


Figure 31: Temperature when changing the distance between NoC and buck converter. (a) Block diagram, (b) simulation results.

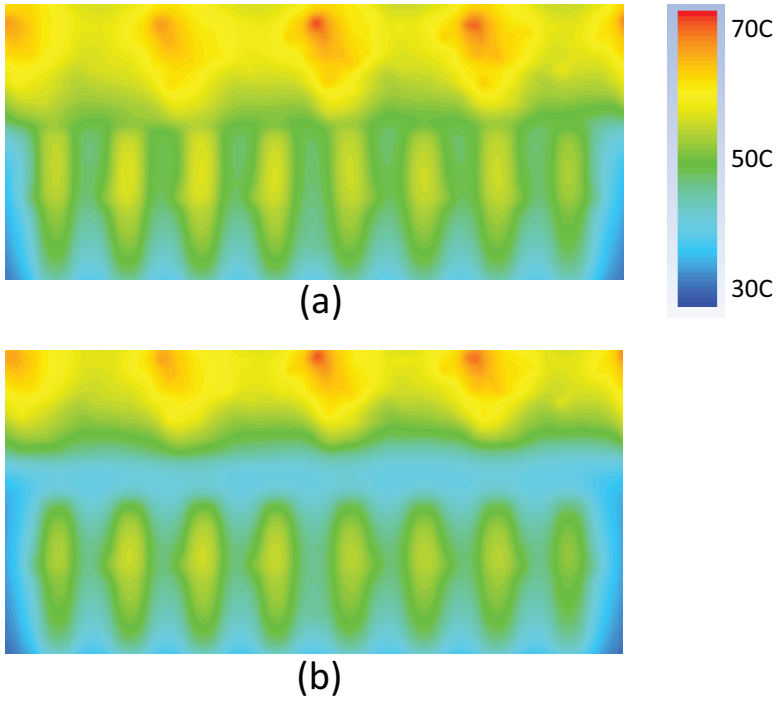


Figure 32: Temperature map when the distance between NoC and the buck converter is (a) 0um (max temp = 71.02°C), (b) 100um, 70.82°C

and the overlapping distance have changed from 0um to 1400um. From Figure 33 (b), 26.35°C can be reduced by avoiding an overlap between the NoC and the inductor. To

reduce the thermal impact between inductor and the NoC, these two components must be far from each other. The temperature map of the simulations done is shown in Figure 34.

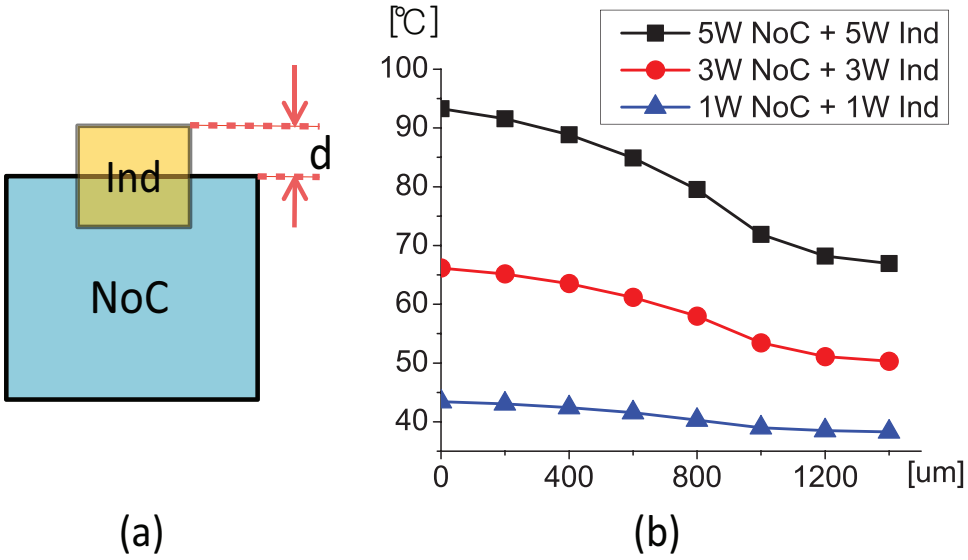


Figure 33: Temperature when changing the overlap distance between NoC and the power inductor. (a) Block diagram, (b) simulation results.

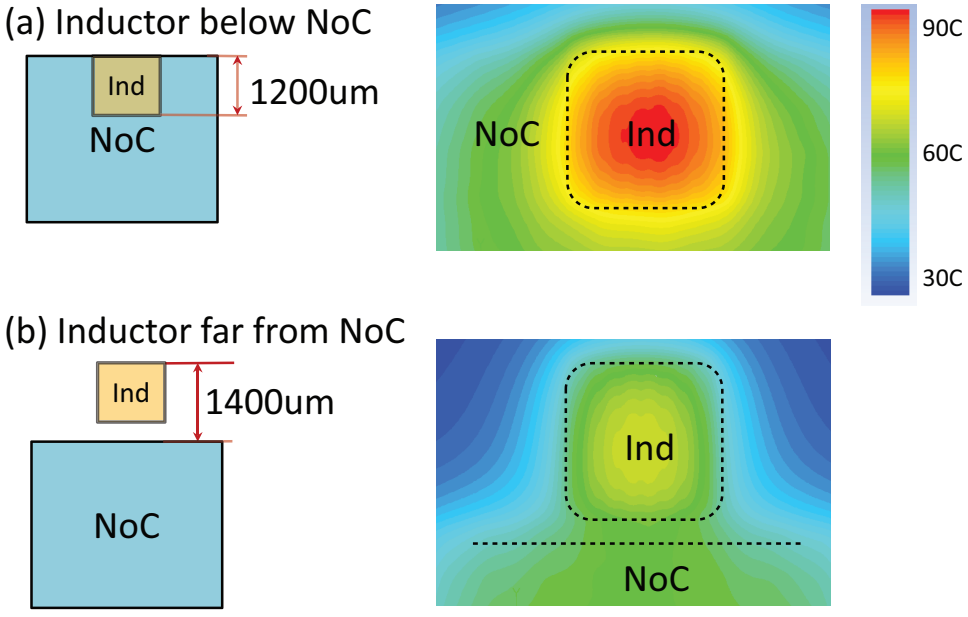


Figure 34: Temperature change when (a) inductor is placed beneath the NoC, (b) inductor is not overlapping the NoC.

2.7 Design Optimization of 2.5D Integrated Voltage Regulator

This section proposes design optimization techniques for temperature reduction.

2.7.1 Design Block Relocation

This study discovered that the placement of the inductor inside the silicon interposer is an important factor to reduce the temperature in the IVR system. Based on this fact, design block relocation is a key technique to reduce the temperature of the whole system. This section assumes that the IVR chip design is fixed and the power inductor design in the silicon interposer is allowed to be modified. As shown in Figure 35, by changing the location of the inductor, a huge temperature reduction is obtained. The inductor cannot be placed on the periphery of the silicon interposer and should be placed beneath the chip. In addition, because of the routing issues, the inductor should be on the surface of the chip. Considering all these factors, instead of placing the inductor in the middle of the chip, it is placed on the bottom right corner to avoid overlap with the NoC and the buck converter as much as possible. The temperature results of this placement with varied power consumptions are shown in Figure 35 (b), and the temperature maps with different inductor placement are shown in Figure 36. By minimizing the overlap between these design blocks, a maximum of 18.41°C temperature reduction is obtained by placing the inductor at a better spot. If the design can fully avoid the overlap between the inductor and the other design blocks, the temperature will decrease further.

2.7.2 Inductor Spreading in Silicon Interposer

In the IVR design, one set of eight coupled power inductors is in the silicon interposer. Assuming that the location of the set of inductors is fixed, yet the design of inductors can be modified, spreading the inductors to a larger footprint is an effective method to reduce temperature. Therefore, this study proposes the inductor spreading technique, by spreading one set of eight inductors into two sets of four coupled inductor or four sets

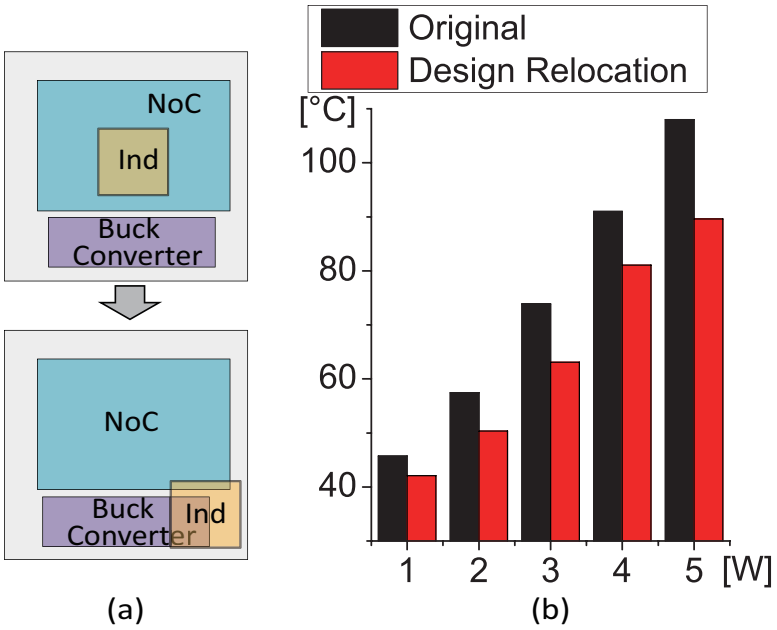


Figure 35: Proposed design block relocation technique. (a) Inductor relocation to minimize the overlap, (b) design block relocation results.

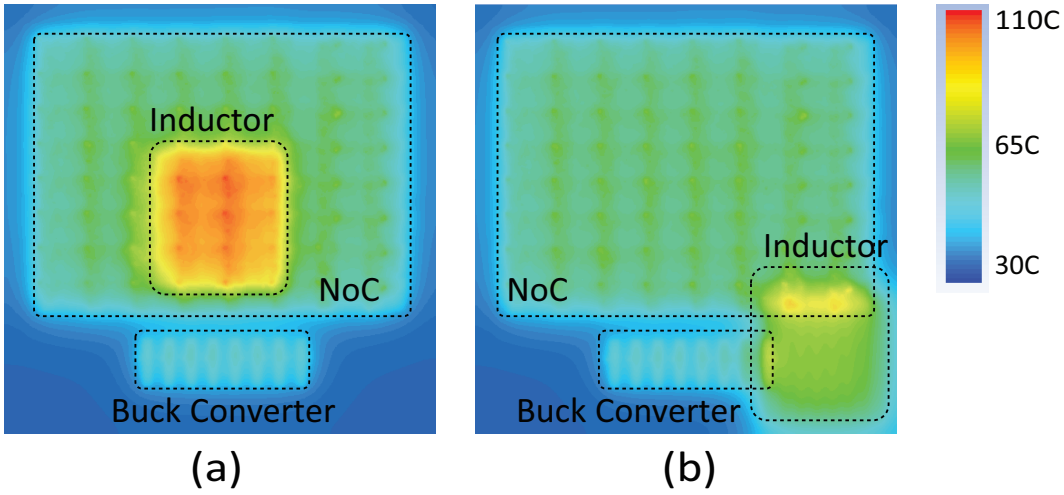


Figure 36: Thermal map of the IVR. (a) Inductor placed in the middle of the chip, (b) inductor placed on the bottom right of the chip to reduce thermal coupling.

of two coupled inductor. In Figure 38 (b) and (c), one set of inductor is split into two and four sets of inductors. The distance between the inductor sets are 200μm, and the maximum temperature is measured when each inductor is operating separately (Figure 38 (a)-(c)) and operating in the full chip (Figure 38 (d)-(f)). Figure 38 shows the thermal map of inductor spreading, and Figure 37 shows the temperature reduction with varied power

consumptions. By spreading the inductor from one set to four sets, maximum 9.73°C in temperature can be reduced. When simulated with the full IVR system, even though the inductors are placed beneath the NoC which consumes high power and thermal coupling is inevitable, maximum of 12.27°C in temperature is reduced by spreading the inductors.

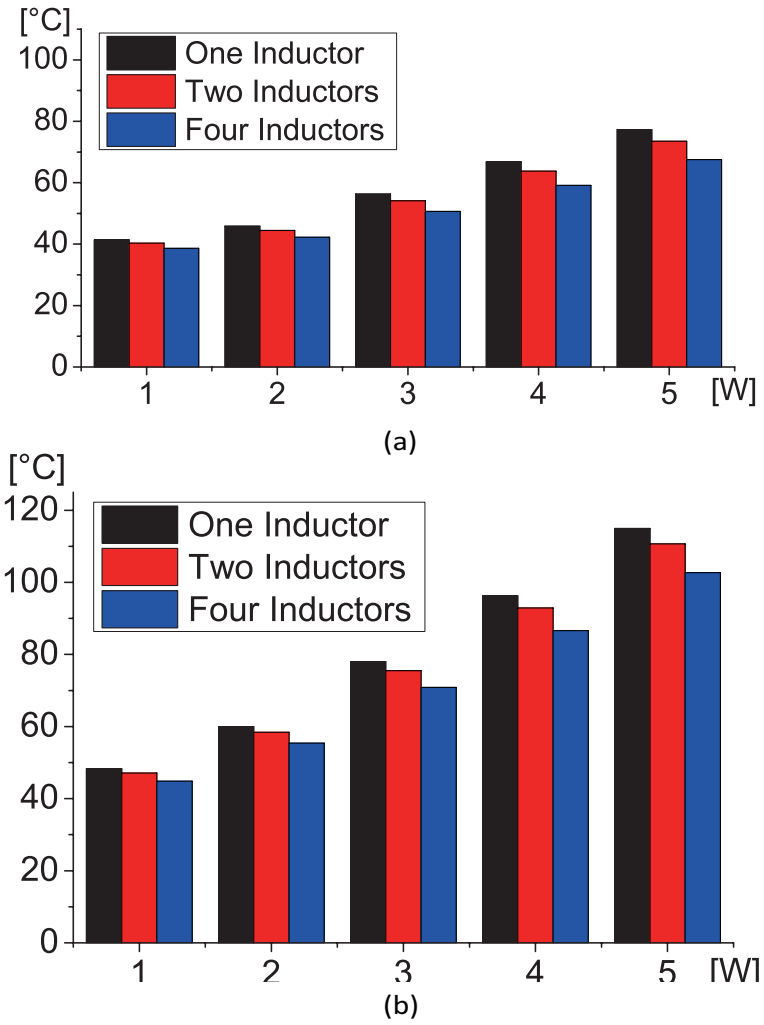


Figure 37: Inductor spreading results: (a) temperature of each inductors, (b) full-chip temperature of the IVR using different inductors.

2.8 Summary

This chapter proposed two co-analysis methodologies for chip, package (silicon interposer), and PCBs. The first study analyzed the severity of IR-drop noise in silicon interposer, and proposed a methodology that can co-simulate IR-drop noise in the entire system.

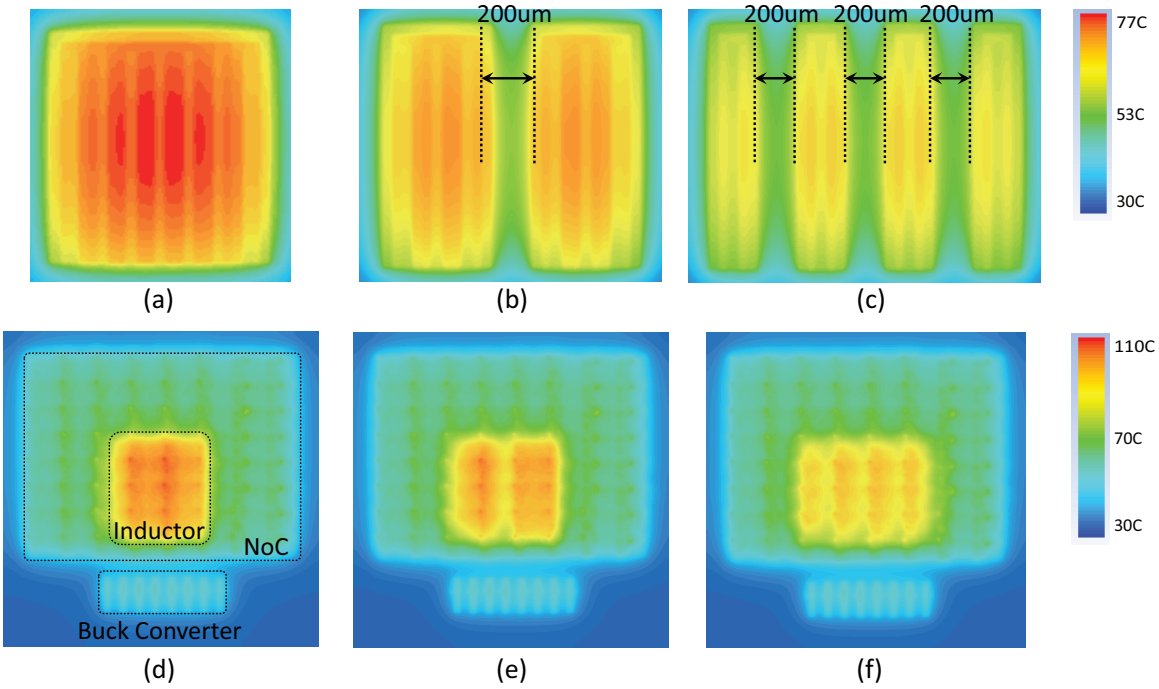


Figure 38: Temperature map of inductor spreading: (a-c) temperature map of inductor with no other heat sources, (d-f) temperature map of the full chip. (a,d) one set of eight coupled inductor, (b,e) two sets of four coupled inductor, (c,f) four sets of two coupled inductor.

This co-simulation methodology can not only simulate 2D IC, package, and PCB, but also simulate a system that consists of 3D IC, silicon interposer, and PCB simultaneously with full transistor level power information. The first study shows that the IR-drop noise in silicon interposer goes up to a few tens of mV, which is more than 8 times organic packages. This study also found that the traditional (= separate) analysis overestimates the IR-drop noise significantly and that the proposed co-analysis provides more accurate power noise values.

The second study proposed a design methodology of performing thermal analysis of analog/digital mixed system in GDSII level and described how this methodology can be utilized to perform thermal analysis on a 2.5D IVR with a silicon interposer. Using the proposed design methodology, this study identified that the power inductor inside the silicon interposer is the hottest component in the system, and the temperature of the inductor alone rises up to 77.3$^{\circ}C$. When the IVR generates $5W$, the maximum temperature rises up

to 114.96$^{\circ}C$ because the power inductor inside the silicon interposer and the NoC on the chip overlaps.

In summary, this chapter showed how co-analysis can successfully be done in two analysis domains: IR-drop and thermal analysis. Despite its challenges, co-analysis is required in many analysis domains for various purposes including IR-drop analysis and thermal analysis. By proper modeling and managing the analysis granularity, this chapter showed the possibility of how co-analysis problem can be tackled when advanced technologies, 3D ICs and silicon interposers, are used in system level.

CHAPTER III

FULL-CHIP SIGNAL INTEGRITY ANALYSIS AND OPTIMIZATION OF 3D ICS

Through-silicon-via (TSV) and three-dimensional integrated circuits (3D ICs) are expected to be the key technology trend in high performance and low power systems [1]. In 3D ICs, dies are stacked vertically, and transistors in different dies are connected by TSVs. TSVs are smaller than off-chip wires, thereby enabling ultra-wide bandwidth and high-speed communication between dies. Industries have started designing 3D DRAMs using TSVs [36], and academia are reporting the impact of TSVs on 3D ICs in many studies [33, 59].

One of the essential signal integrity (SI) characteristics in studying TSVs is coupling. In 2D ICs, metal-to-metal is the main source of noise coupling. Two adjacent metal wires form a parallel capacitor, and noise voltage travels from an aggressor to a victim through close metal wires (capacitive coupling). However, two adjacent TSVs form a complex coupling network due to its surroundings in 3D ICs. TSV-to-TSV coupling not only forms a capacitive coupling network, but it also forms other complex coupling networks. These coupling networks cause significant coupling noise between two adjacent TSVs. Therefore, a signal path that includes TSVs can suffer from significant noise in 3D ICs.

Authors of [11, 83, 14] showed S-parameter-based coupling analysis assuming that all ports are under 50-Ω termination. However, it is not possible nor practical to create 50-Ω termination inside an IC. Therefore, this chapter first applies a lumped circuit model with a realistic high-impedance termination condition to analyze TSV-to-TSV coupling in 3D ICs. The results show that the proposed circuit-model-based analysis is highly accurate and the difference between different termination conditions is huge. Then, this chapter studies the multiple TSV-to-TSV coupling effect inside 3D ICs on a full-chip level. The

true phenomena that take place inside the ICs are described and a compact model that captures the coupling effect between multiple TSVs is proposed. Then, a methodology that performs an analysis of multiple TSV coupling on a full-chip level is proposed. Based on the proposed methodology, this chapter also studies what the critical factors are that affect noise coupling and delay in 3D ICs.

3.1 Electrical Model of TSVs

TSV-to-TSV crosstalk analysis requires electrical models for a physical structure that consists of TSVs, insulator, silicon substrate, bumps, and I/O drivers. Thus, this section shows the physical structure and its electrical model of a 3D IC channel. Then, this research validates the component models using a commercial simulator. Figure 39 shows a simplified model of a TSV channel, and Figure 40 shows its equivalent lumped circuit model. The TSV at the right hand side is the aggressor, which is driven by Port1. The TSV to the left is the victim.

The lumped circuit modeling can be used because the elements this study is modeling are smaller than 100um, which are all shorter than the $1/20\lambda$ wavelength of 20GHz. The electrical parameters and process technology nodes used in this model are presented in Table 4. Since TSVs are made of conducting material such as copper or tungsten, a TSV is modeled as a series connection of a resistor (R_{TSV}) and an inductor (L_{TSV}). Silicon dioxide insulator between TSV and silicon substrate is modeled as a capacitor (C_{TSV}). On the other hand, silicon substrate can be modeled as a capacitor (C_{si}) in parallel with a resistor (R_{si}) as shown in Figure 40. Mutual inductance exists between two TSVs, which also has to be modeled ($M_{\mathrm{TSV-TSV}}$). In order to compute the capacitances and the resistances, this study uses the following equations presented in [83]:

$$C_{\mathrm{TSV}} = \frac{1}{4} \frac{2\pi\epsilon_0\epsilon_{\mathrm{r}}}{\ln\left(\frac{r_{\mathrm{TSV}}+t_{\mathrm{ox}}}{r_{\mathrm{TSV}}}\right)} \cdot l_{\mathrm{TSV}} \tag{5}$$

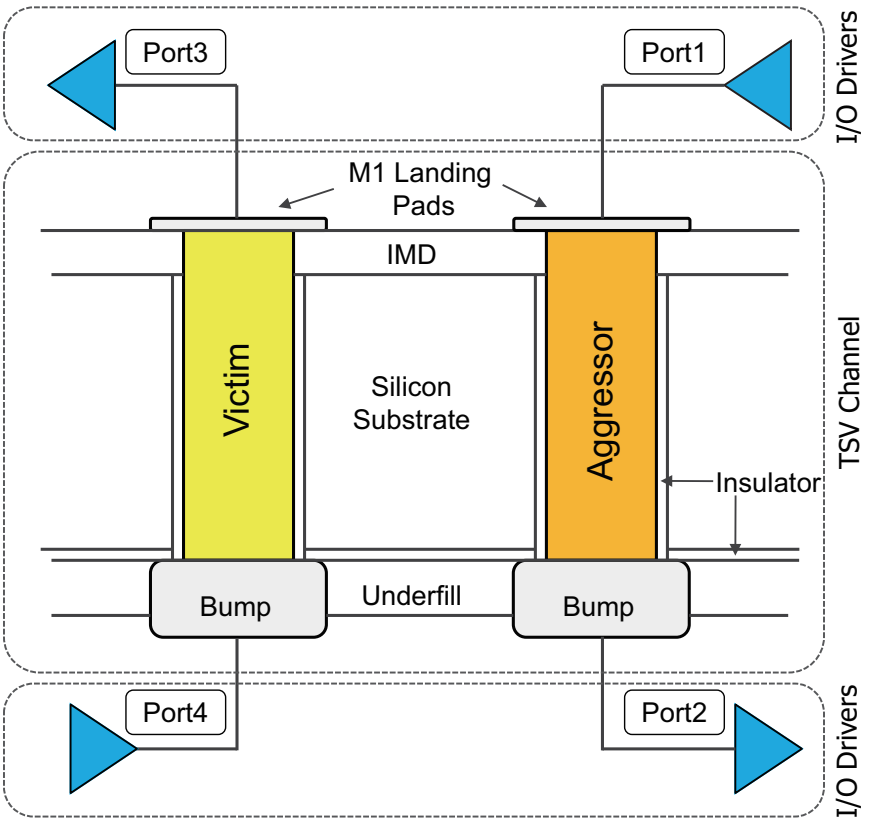


Figure 39: A simplified model of TSVs and I/Os in 3D IC.

Table 4: Electrical parameters used in this paper

Parameter	Value
TSV diameter	$2.5\mu m$
TSV height	$75\mu m$
Insulator thickness	$0.5\mu m$
Bump pad diameter	$5.0\mu m$
Bump height	$10\mu m$
Dielectric constant of liner	4
Dielectric constant of underfill	4
Process technology	Nangate $45nm$
Supply voltage	$1.2V$

$$C_{\text{si}} = \epsilon_0 \epsilon_{\text{si}} \frac{2(r_{\text{TSV}} + t_{\text{ox}}) + \alpha}{d} \cdot l_{\text{TSV}} \tag{6}$$

$$R_{\text{si}} = \rho_{\text{si}} \cdot \frac{d}{2(r_{\text{TSV}} + t_{\text{ox}}) + \alpha} \cdot \frac{1}{l_{\text{TSV}}} \tag{7}$$

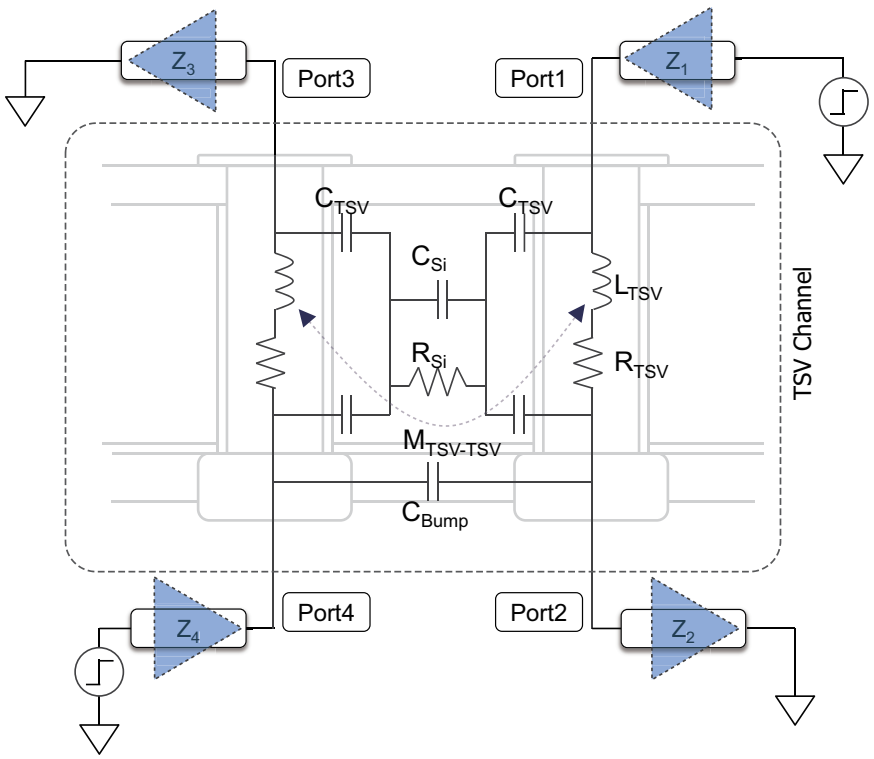


Figure 40: Equivalent lumped circuit model for the TSV channel.

$$C_{\mathrm{Bump}} = \frac{\epsilon_0 \epsilon_{\mathrm{r}}}{d - 2r_{\mathrm{Bump}}} \cdot \pi \cdot r_{\mathrm{Bump}} \cdot l_{\mathrm{Bump}} \tag{8}$$

where r_{TSV} is the TSV radius, l_{TSV} is the TSV height, t_{ox} is the thickness of the insulator, d is the pitch between two TSVs, r_{Bump} is the radius of a bump, and l_{Bump} is the height of a bump.

Regarding Equation 6 and 7, many papers [83, 8] have mentioned this as the electromagnetic formula of capacitance between two parallel pillars. This may be effective in cases where no other TSVs are interfering inside the fields that are generated between the two pillars. However, in such cases this assumption may not be valid. Therefore this study proposes a formula regarding the silicon substrate as a parallel plate capacitor that considers the effective volume of the silicon substrate between two TSVs. In Equation 6 and 7, this study uses scaling factors (α) that has the value of $24\mu m$.

Figure 41 shows the coupling coefficients (s_{31}) obtained from a commercial 3D electromagnetic simulator (Ansys HFSS) and the proposed lumped circuit model when the distance between two TSVs is $10\mu m$. Note that the 3D simulator can only support termination condition of 50-Ω. As the figure shows, the proposed TSV model is very accurate and the maximum difference is less than 1 dB.

In the basis of the proposed TSV model, this study uses $1\times$ inverter (w_{p}=$260nm$, w_{n}=$130nm$) for each I/O driver. A driver is modeled as a resistance (output resistance) connected to the supply voltage, and a load as a capacitance (input capacitance) connected to the ground. Here, this study shows the voltage noise level observed at Port3 when 1GHz digital signal is inserted at port 1 in Figure 42. Despite the small driver size on port 1, the peak noise is $-101.7mV$, which is not negligible.

3.2 Analysis of TSV-to-TSV Crosstalk

3.2.1 Crosstalk Equations Under High-Impedance Termination

In the frequency range under 20GHz, silicon substrate, bumps and the insulator (silicon dioxide) around TSVs form a channel having very high impedance. On the other hand, the impedance composed of TSV resistance and inductance is very low. If low impedance components can be ignored, the lumped circuit model in Figure 40 can be simplified as a model having only high-impedance components as shown in Figure 43. Applying Kirchhoff's laws to the model in Figure 43 (b), the following matrix for V_1 and V_2 can be obtained:

$$\begin{bmatrix} \frac{1}{Z_1}+\frac{1}{Z_2}+\frac{1}{Z_5} & -\frac{1}{Z_5} \\ -1 & 1+\frac{Z_5}{Z_3}+\frac{Z_5}{Z_4} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} \frac{V_{\mathrm{in}}}{Z_1} \\ 0 \end{bmatrix} \tag{9}$$

where Z_5 is the impedance of the TSV channel in the simplified model. Solving for V_2, the following equation can be finally obtained:

$$V_2 = V_{\mathrm{in}} \cdot \frac{Z_2 Z_3 Z_4}{Z_1 \cdot Z_A + Z_2 Z_3 Z_4 + Z_5 \cdot Z_B} \tag{10}$$

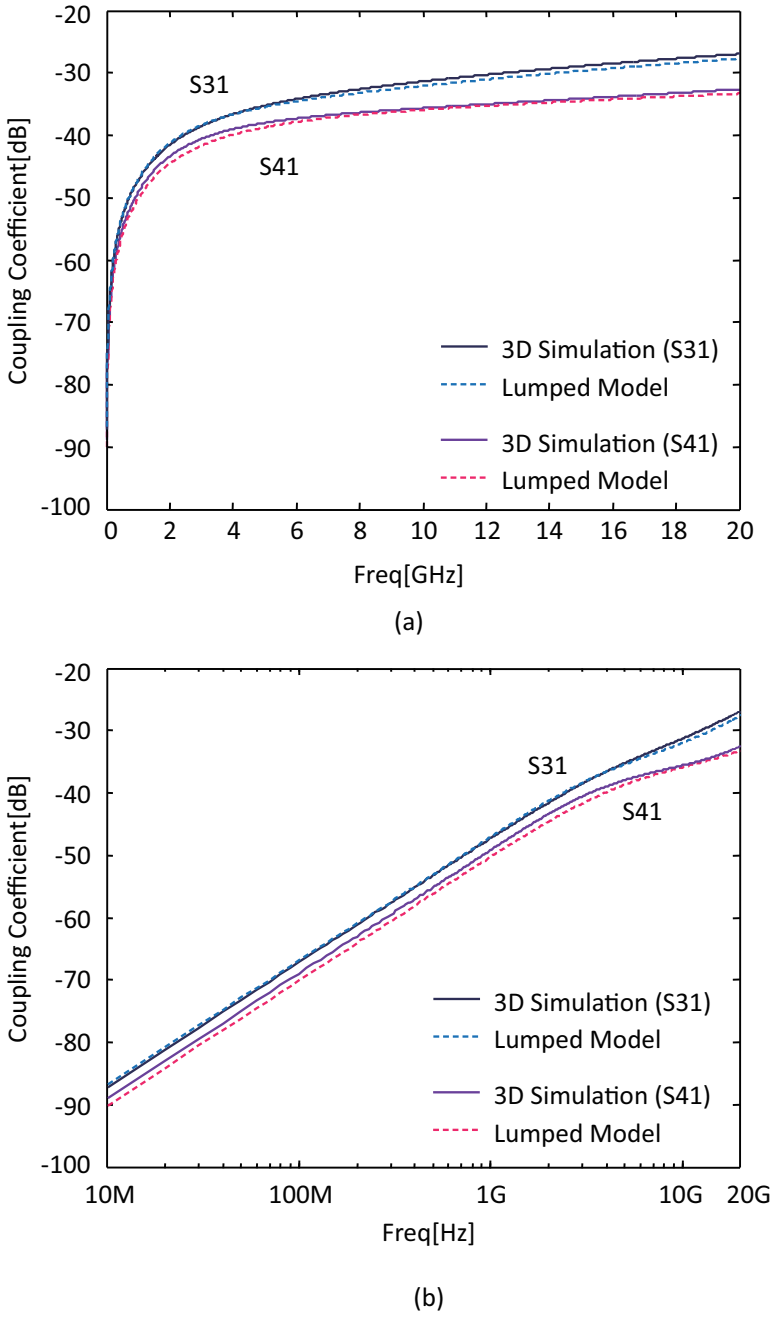


Figure 41: Coupling coefficients obtained from a 3D simulator model and the proposed lumped circuit model when the TSV-to-TSV distance is $10\mu m$. (a) Linear scale, (b) Log Scale

where

$$Z_A = Z_2Z_3 + Z_2Z_4 + Z_3Z_4 + Z_3Z_5 \quad (11)$$

$$Z_B = Z_1Z_4 + Z_2Z_3 + Z_2Z_4 \quad (12)$$

$$Z_5 = \frac{Z_{C_{\text{Bump}}}(Z_{C_{si}}//Z_{R_{\text{si}}} + 2Z_{C_{\text{TSV}}})}{Z_{C_{\text{si}}}//Z_{R_{\text{si}}} + Z_{C_{\text{Bump}}} + 2Z_{C_{\text{TSV}}}} \quad (13)$$

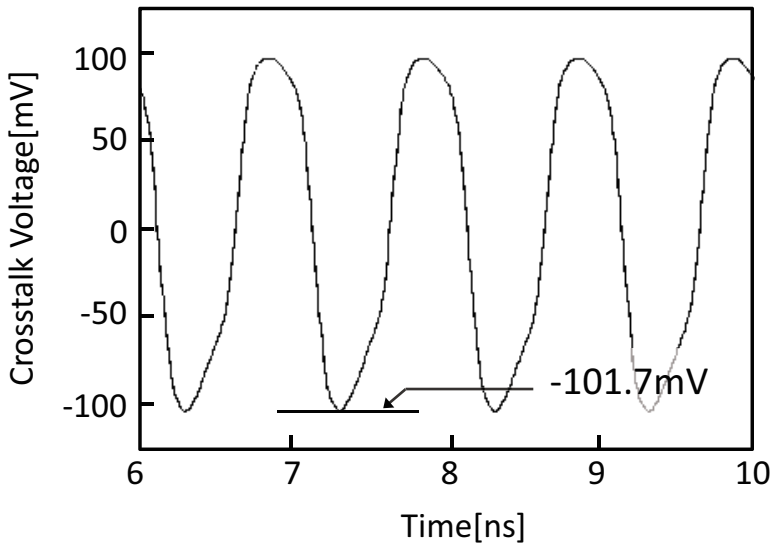


Figure 42: Crosstalk voltage observed at port3 when 1.2V, 1GHz digital signal inserted to port1. ($1\times$ driver, TSV-to-TSV distance: $10\mu m$.)

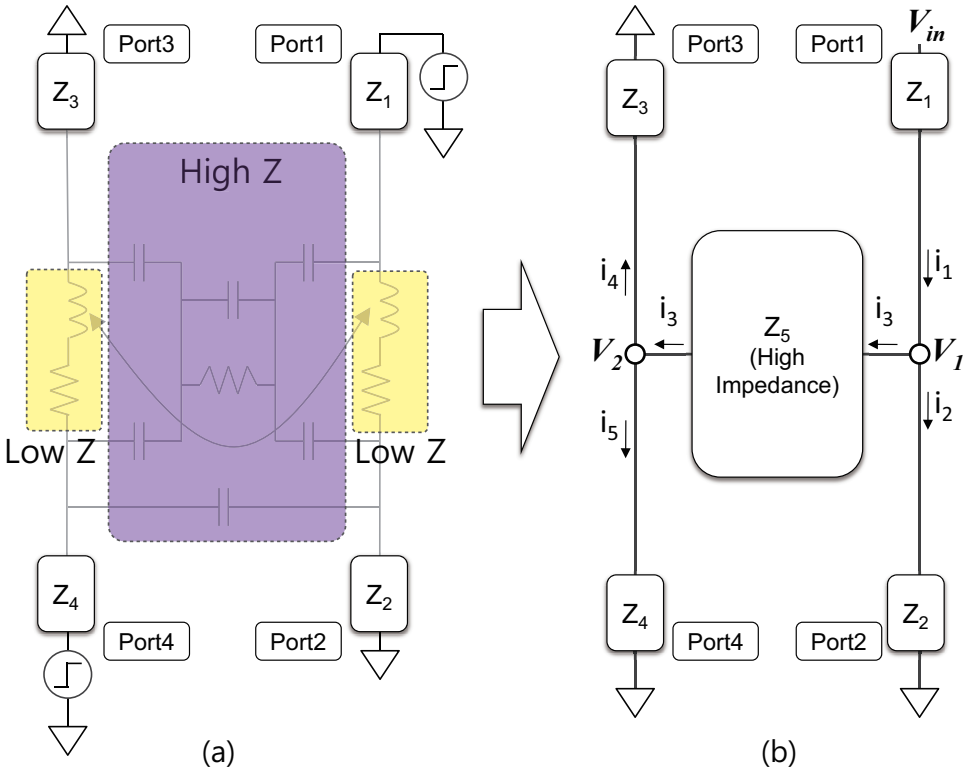


Figure 43: (a) Impedance level of each component in the lumped circuit model, (b) Simplified model for coupling analysis.

Equation 10 shows that the coupling between two TSVs depends not only on the channel impedance (Z_5) between the TSVs, but also on the termination condition (Z_2, Z_3, Z_4)

and the driver condition (Z_1).

3.2.2 Comparison of Termination Conditions

S-parameter-based coupling analysis assumes a 50-Ω termination condition, which is very unlikely inside an IC. Therefore this study changes termination conditions, compute coupling, and compare their results in this section.

Figure 44 compares two different termination conditions. When all ports are terminated with 50-Ω resistance (solid line), the coupling coefficient is below -30dB even in the highest frequency region (under 20GHz). However, when all ports are terminated in high impedance ($1\times$ driver at all ports), the coupling coefficient reaches almost up to -10dB. The coupling coefficient in this case is so high that it causes serious crosstalk in over GHz high frequency range. This cannot be observed if 50-Ω termination is assumed, and this is the special channel-termination condition in 3D ICs.

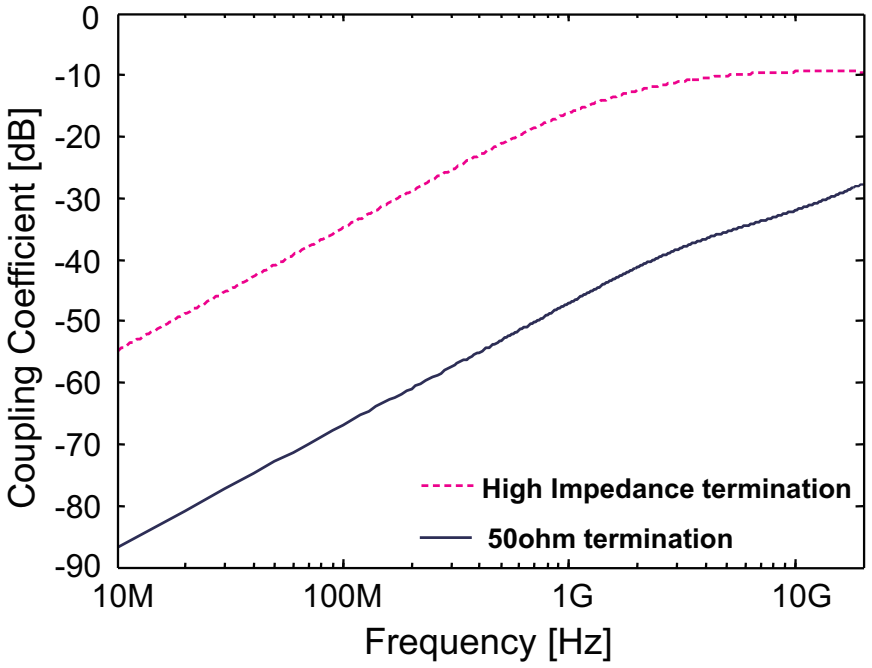


Figure 44: Coupling coefficients of the 50Ω termination condition (solid line) and the high impedance termination ($1\times$ driver, dotted line) condition.

3.2.3 Macro Impact of Port Impedance on TSV-to-TSV Coupling

This section explains the macro impact of port impedance on TSV-to-TSV coupling using Equation 10. Regarding Equation 10, it is Z_2, Z_3, and Z_4, which are the dominating variables inside the total equation. Assume a typical signal coupling channel where the ports are driven by a typical size driver (Z_1). In this case the port impedances at Z_2, Z_3, Z_4 are in the same scale. Replacing Z_2, Z_3, and Z_4 with the same term Z_{port} ($Z_2 = Z_3 = Z_4 = Z_{port}$), Equation 10 can be rewritten as:

$$V_2 = V_{\text{in}} \cdot \frac{{Z_{\text{port}}}^3}{{Z_{\text{port}}}^3 + {Z_{\text{port}}}^2(3Z_1 + 2Z_5) + 2Z_{\text{port}}Z_1Z_5} \tag{14}$$

Equation 14 shows that if the port impedance is much higher than the channel impedance, the coupling level can be very large, even close to the agrressor voltage.

However, there are factors limiting the coupling voltage to a certain range. In the previous analysis, the TSV capacitance at the other side (not on the coupling side) that connects to ground through substrate was not considered. This capacitance is large, and in parallel with the port impedance. A TSV not only sees the port impedance but also sees the GND capacitance. With this capacitance, the coupling voltage is limited to a certain level. Thus, even when a port impedance is too high, the GND capacitance acts like a buffer and screens out the high port impedance (see Figure 45).

3D ICs deal with a situation where the ports' impedance (less than a few fF capacitance) is much higher than the coupling channel impedance (tens of fF capacitance series to fF capacitance and kΩ resistance). However, due to the high capacitance between the TSVs and the GND, this capacitance limits the coupling voltage to be at a certain level.

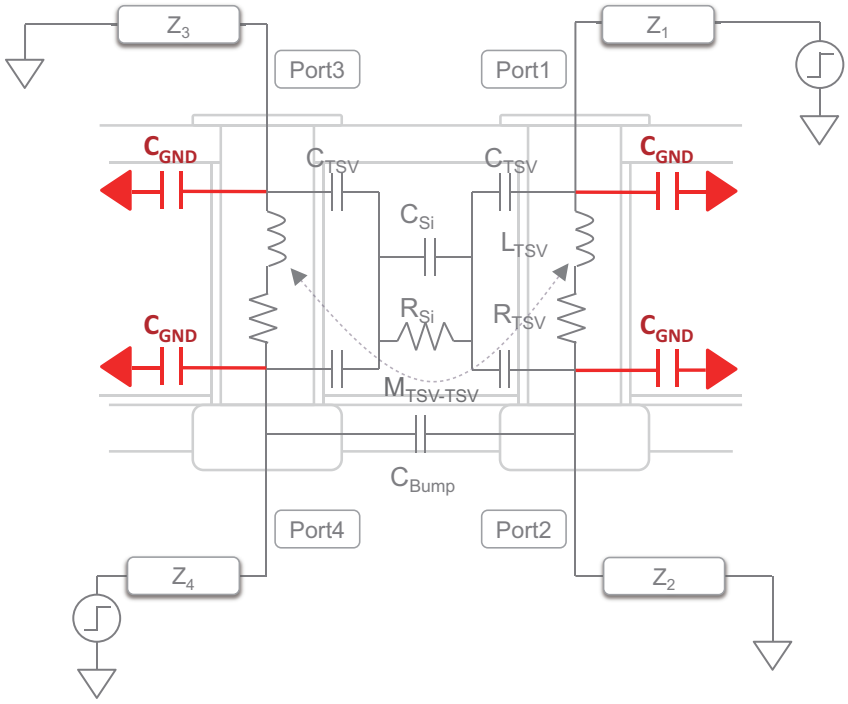


Figure 45: Impact of GND capacitance in TSV coupling channel.

3.2.4 Micro Impact of Port Impedance on TSV-to-TSV Coupling

This section explains the micro impact of port impedance on TSV-to-TSV coupling when all port impedance numbers are not the same, but are fixed in a specific range, using Equation 10 and Figure 46. Here, the individual role of each ports to channel coupling is discussed. First, when the driver (Port1) is big (low output resistance = low Z_1), it becomes a strong aggressor, and increases crosstalk. This is also observed quantitatively in Equation 10 because Z_1 exists only in the denominator. On the other hand, if the sink (Port2) is big (high input capacitance = low Z_2), the impedance at Port2 becomes low and the impact of the aggressor decreases. Similarly, if the sink (Port3 or Port4) in the victim net is big (high load capacitance = low Z_3, high load capacitance and low output resistance = low Z_4), it reduces the crosstalk.

In fact, Equation 10 can be rewritten as:

$$V_2 = V_{\text{in}} \cdot \frac{ax}{(a+b)x+c} = \frac{a}{a+b} \cdot \left(1 - \frac{c}{(a+b)x+c}\right) \tag{15}$$

where x is a variable which can be one of Z_2, Z_3, or Z_4 (say, x=Z_2), and a, b and c are constants when the frequency is fixed and Z_n ($\neq$ x) is a constant (say, Z_3 and Z_4 are fixed). This equation transformation describes that V_2 increases monotonically as x increases. Therefore, high load capacitance (low load impedance) reduces the impact of the aggressor. A stronger driver at victim net and a weaker driver at aggressor net also reduce the coupling level.

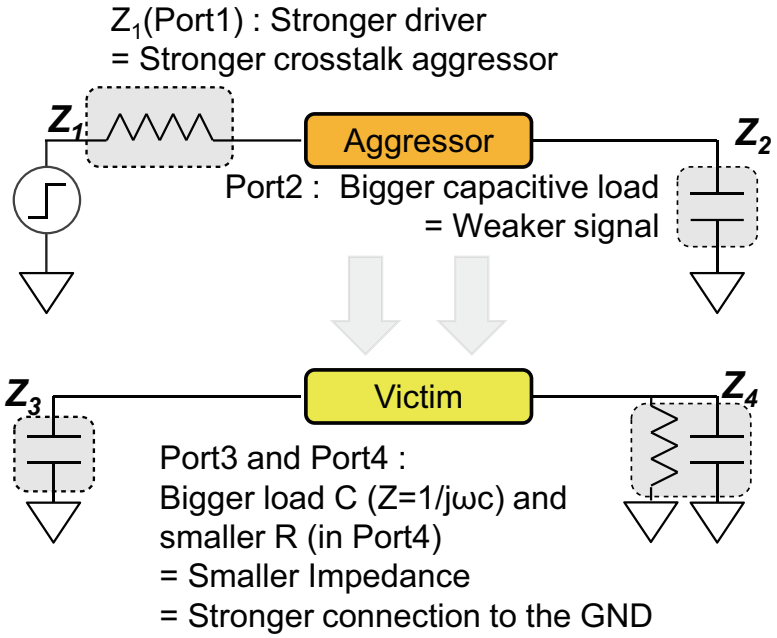


Figure 46: Visualization of a driver strength, load impedance, and the relationship between the aggressor and the victim.

3.2.5 Dependency of Channel Impedance on Low Frequency

Unlike wire coupling channels, TSV coupling has a very unique coupling channel characteristic. Due to the various types of components in the coupling channel, the impedance of the channel differs in each frequency range (see Figure 47). Thus, by analyzing how the lumped components react to each other in the specific frequency range, how coupling would occur in each frequencies can be predicted. These frequencies can be categorized in to three regions: the low frequency region ($< 1GHz$, (I)), the middle frequency region ($1GHz$ to $8GHz$, (II)), and the high frequency region ($> 8GHz$, (III)). Here, C_{bump} is

ignored in the analysis due to the high impedance in all frequency regions, and $M_{TSV-TSV}$ is also ignored due to the small impact it has in the analysis.

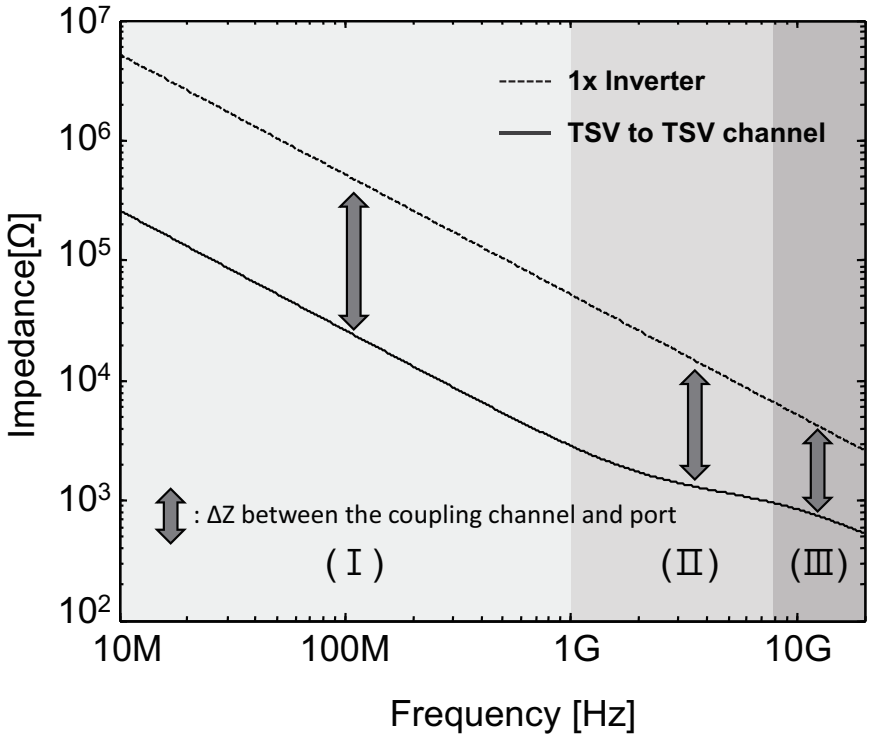


Figure 47: Impedance difference between the silicon substrate channel, and the gate capacitance in different regions: (I) low frequency ($<$ 1GHz), (II) middle frequency (1GHz to 8GHz), (III) high frequency ($>$ 8GHz).

In the low frequency region, the coupling path can be defined by C_{TSV} and R_{si}. Since the impedance of C_{si} is very high in this region, all the coupling current will detour through R_{si} (see Figure 48). Thus, inside the silicon substrate, the dominant coupling factor is the resistive coupling by R_{si}. The impedance of the channel in this frequency will be the impedance sum of C_{TSV} and R_{si}. However, since $Z_{R_{si}}$ is very low compared to $Z_{C_{TSV}}$, the impedance of the channel in this frequency can be expressed as the impedance sum of C_{TSV}s.

$$Z_{\text{Channel,Lowfreq}} \approx Z_{\text{C}_{\text{TSV}}} \tag{16}$$

On the contrary to the common belief, this phenomena describes that changing the

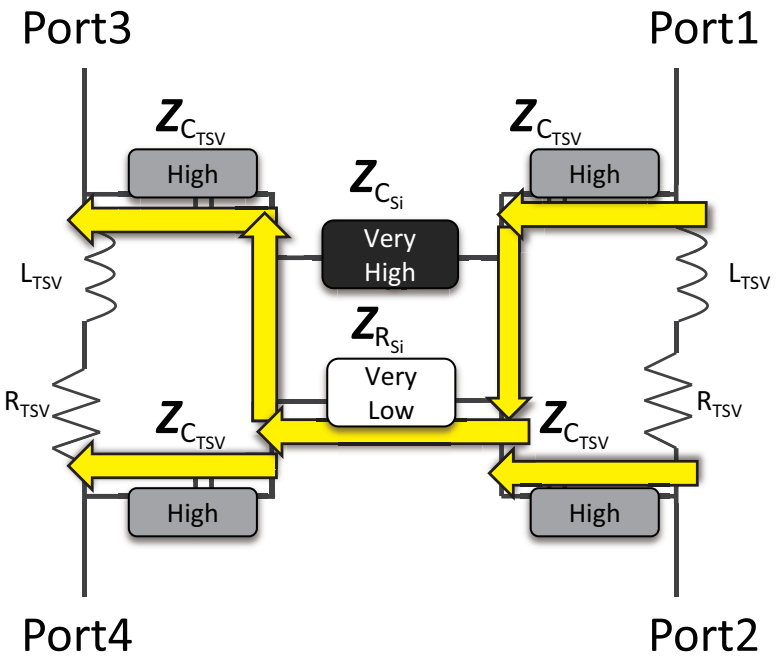


Figure 48: Coupling path in the low frequency region.

distance between TSVs to alleviate coupling in this frequency region does not work well. For a digital signal in a specific frequency, it has its harmonic components (up to $7\times$). However, if there is a digital signal, whose frequency harmonics are all inside this low frequency range, changing the distance between TSVs would not have a significant impact on alleviating coupling.

There are two reasons for this: First, regarding Figure 47, the difference between channel impedance and the port impedance in the low frequency region is very big (more than 20dB). Due to the huge difference of the impedance, a slight change in the channel impedance would not result in a big difference on the total coupling (see equation 6). The other reason is that the channel impedance is mainly determined by the TSV capacitance(see equation 12). C_{TSV} is defined as the capacitance between TSV and silicon substrate, which is mainly determined by the thickness of insulator. Therefore C_{TSV} is a fixed value once a 3D IC is made, and is insensitive with TSV distance change. Therefore, in this low frequency region, changing the distance between TSVs would not have a big impact. As it can be seen in Figure 49, the crosstalk voltage of a 100MHz digital signal

that is generated at port 1 (size of 1×) in 10um distance TSVs and 30um TSVs are almost the same.

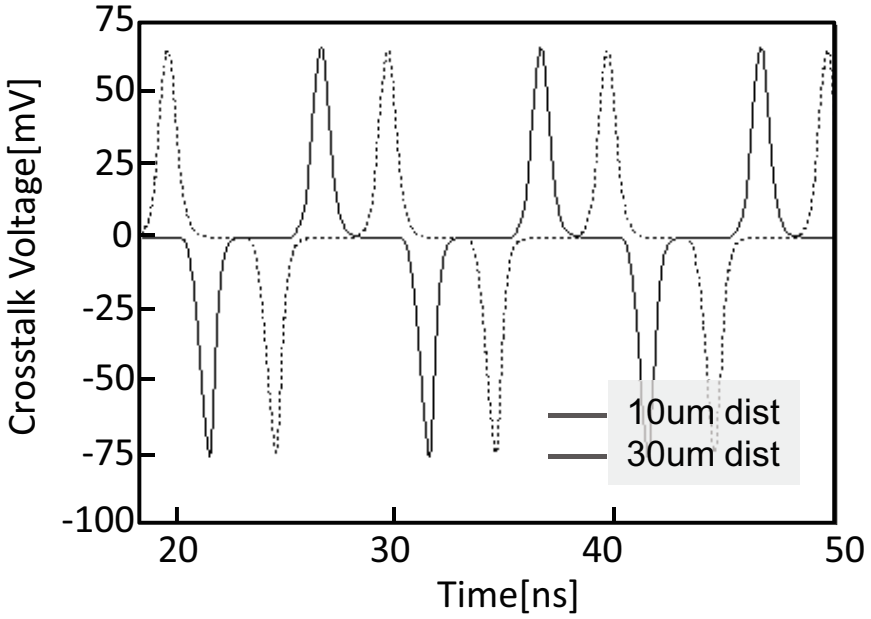


Figure 49: Crosstalk voltage of 100MHz digital signal when the distance between TSV is 10um, and 30um (1× driver).

3.2.6 Dependency of Channel Impedance on Middle Frequency

In the middle frequency region ($1GHz$ to $8GHz$), the impedance of C_{TSV} becomes sufficiently low, and the impedance of C_{si} becomes comparable with $Z_{R_{si}}$. However, the impedance of C_{si} is still higher than R_{si} in this region, and most of the coupling current flows through R_{si}. The new phenomena that is observed in this region is that due to the smaller difference of these two impedances, C_{si} becomes a path for the coupling current to flow (see Figure 50). In this region, neither R_{si} nor C_{si} is a dominant coupling factor inside the silicon substrate. Both R_{si} and C_{si} affects the substrate coupling.

In summary, In the middle frequency region, impedance of all the components become similar to each other. Unlike in the low frequency region, no component has the dominating impedance value, and all the components are equally responsible for the coupling path. Thus impedance of the channel in this frequency region can be expressed as the sum of all

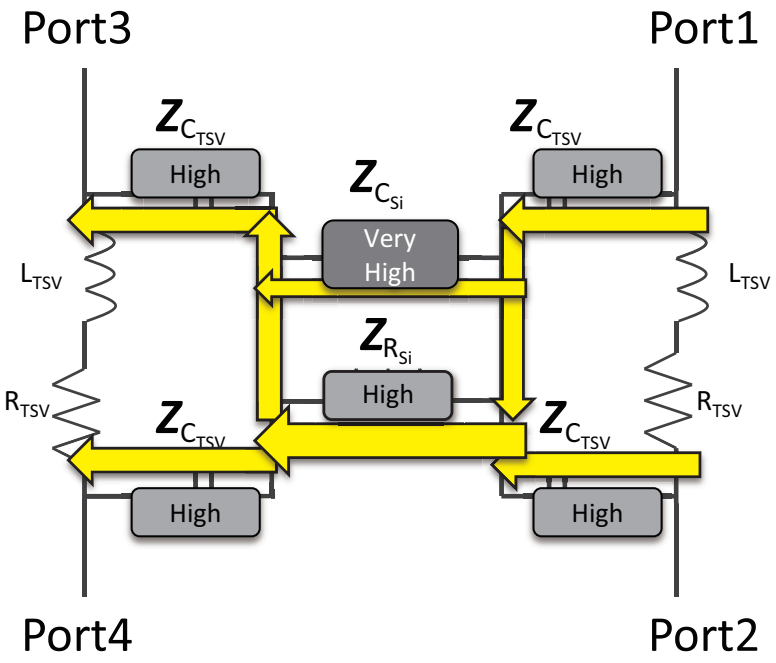


Figure 50: Coupling path in the middle frequency region

components inside the channel.

$$Z_{\text{Channel,Midfreq}} = Z_{\text{C}_{\text{TSV}}} + Z_{\text{C}_{\text{si}}} // Z_{\text{R}_{\text{si}}} \tag{17}$$

In the middle frequency region, the difference of impedance between port and channel becomes smaller (see Figure 47). Now that ΔZ between the channel and port is smaller, the output starts to respond to the change of numbers of each components (R_{si}, C_{si}, C_{TSV}). Starting from this region, the coupling voltage becomes dependant on the TSV distance. Figure 51 describes the effect of TSV-to-TSV coupling in this region. For a signal whose harmonics are partly in the middle frequency region, the change of distance between TSVs affects TSV-to-TSV coupling.

3.2.7 Dependency of Channel Impedance on High Frequency

In the high frequency region (Over $8GHz$), all capacitance components (C_{si}, C_{TSV}) have an impedance lower than the resistance of silicon substrate (R_{si}). Since R_{si} is the highest impedance showing in this region, the coupling current detours R_{si}, and mostly flows

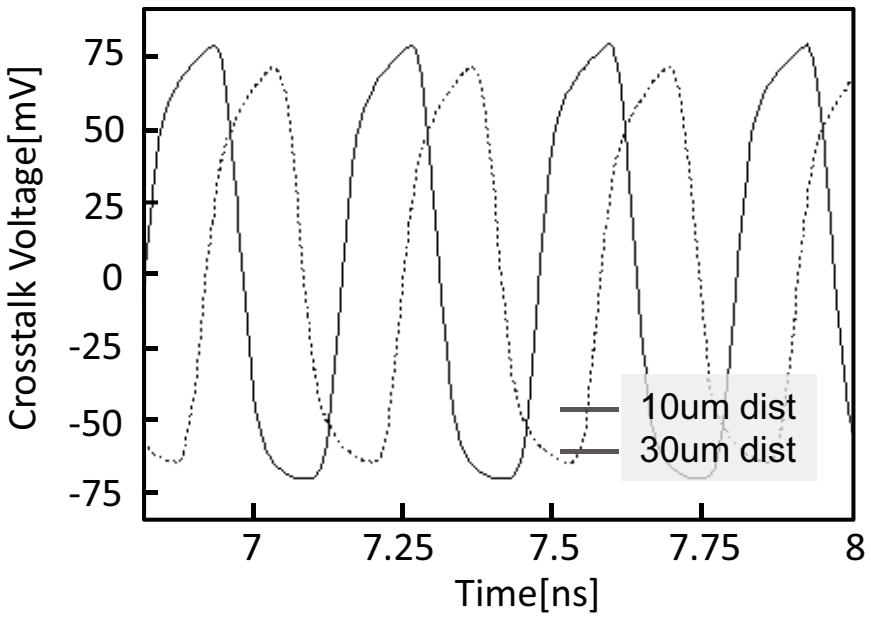


Figure 51: Crosstalk voltage of 3GHz digital signal when the distance between TSV is 10um, and 30um (1× driver).

through Csi inside the substrate (see Figure 52). In this region, the dominant coupling that occurs inside the silicon substrate is capacitive coupling through C_{si}. Since the capacitance of C_{si} is mostly smaller than C_{TSV}, the impedance of C_{si} is bigger than C_{TSV}. Thus, the impedance in this region is dominated by the capacitance of silicon substrate.

$$Z_{\text{Channel,Highfreq}} \approx Z_{\text{C}_{\text{si}}} \tag{18}$$

The region that the coupling voltage is the most sensitive to the change of distance is the high frequency region. C_{si} is a factor that is solely determined by the change of distance. Since the dominating factor of $Z_{Channel,Highfreq}$ is C_{si}, this region is most sensitive to TSV distance.

The overall trend on TSV coupling to the change of distance is described in Figure 53. In the low frequency region (region 1), distance change among TSVs do not result in a big change in the coupling level. This is because the dominant component to the coupling is C_{TSV}, and C_{TSV} hardly changes with TSV distance. In the middle frequency region (region 2), distance change between TSVs starts to change the level of coupling. This is

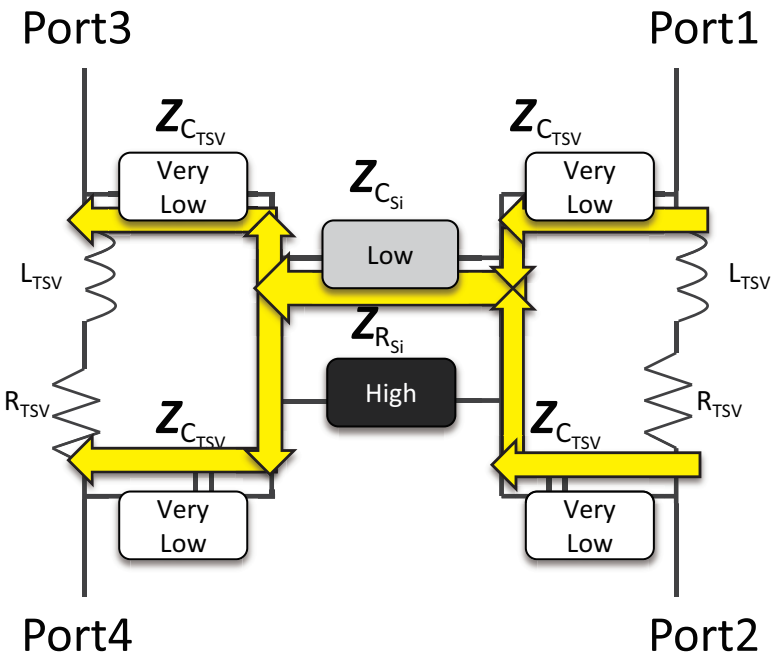


Figure 52: Coupling path in the high frequency region

because C_{si} and R_{si} begin to have impact on $Z_{Channel,Midfreq}$ along with C_{TSV}. Since C_{si} and R_{si} are dependent on TSV distance, the coupling level starts to react on the change of TSV distance. In the high frequency region (region 3), the change in TSV distance has the biggest impact on coupling level. This is due to C_{si}, which dominates the impedance of the coupling channel.

3.2.8 A New Technique for Coupling Reduction

The most conventional way to reduce coupling is to increase the distance between TSVs. However, through this study in the previous sections, it was shown that the change of distance between TSVs may not be very effective in reducing coupling. Therefore, a new technique to reduce coupling between TSVs is proposed.

TSV coupling reduction can be obtained by decreasing gate size of the aggressor, or increasing gate size of all other ports. Figure 54 compares the results of three different simulation settings when port 1 sends 1GHz digital signal ($1\times$ driver). When the load impedance is fixed but TSV-to-TSV distance increases from $10\mu m$ to $30\mu m$, the peak noise

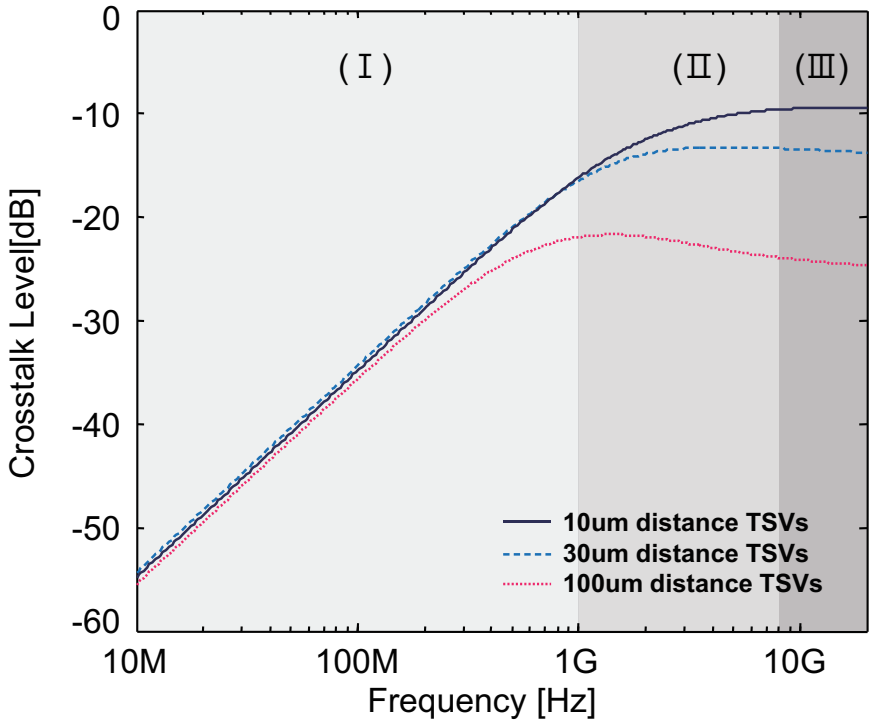


Figure 53: Frequency dependency on TSV coupling to distance on high impedance termination: (I) low frequency, (II) middle frequency, (III) high frequency.

decreases from $101.74mV$ to $95.13mV$. Although the distance change is big ($3\times$), the noise reduction is not as significant as the distance changes. This is because the $1GHz$ signal along with its major harmonics are in the low frequency region, where distance between TSVs does not affect the coupling level. On the other hand, when the TSV-to-TSV distance is fixed ($10\mu m$), while the load becomes $2\times$ bigger, the peak noise decreases from $101.74mV$ to $58.65mV$. Therefore, it is observed that gate sizing (by increasing the gate size at the sink node, or increasing the gate size at the driving node on the victim net.) has more impact on coupling than increasing TSV-to-TSV distances.

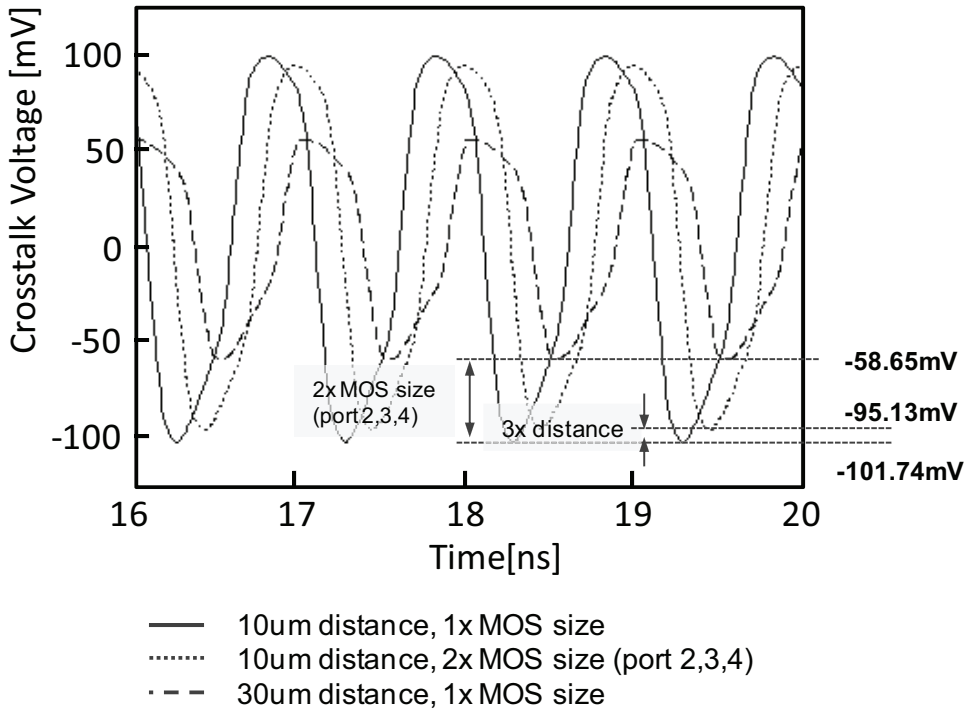


Figure 54: Crosstalk voltage of 1GHz digital signal when distance and gate size have changed.

3.3 Motivation for an Accurate Full-Chip Analysis

This section describes the motivation of an accurate full-chip analysis and show the important findings. This section shows why [45] is inaccurate. In this chapter, the TSVs of a diameter of 5μm, a height of 60μm, a SiO_2 liner of 0.5 μm, and a minimum pitch of 15μm is used.

3.3.1 Maximum Coupling Capacitance

In [45], the authors assumed that silicon substrate capacitance depends on the distance between two TSVs only. However, when a victim TSV is surrounded by more than one aggressor, the total coupling capacitance of the silicon substrate has a maximum limit and does not increase linearly. Many TSV modeling papers [45] [35] claim that the silicon substrate capacitance follows Equation 19, which is the capacitance between two parallel, circular conducting wires,

$$C_{\text{si}} = \frac{\pi \epsilon_0 \epsilon_{\text{si}} L}{\ln[(P/2r) + \sqrt{(P/2r)^2 - 1}]} \tag{19}$$

in which, ϵ_{si}, L, P, and r are the permittivity of the silicon substrate, the height of the TSVs, the pitch between the TSVs, and the radius of the TSVs, respectively. By this equation, when the coupling capacitance between an aggressor and a victim in a certain pitch is 1x, the victim will see 8x coupling capacitance when there are eight aggressors in every direction.

However, Equation 19 is correct only when there are no other neighbors near the two TSVs. When TSV aggressors are close to another aggressor, the total substrate capacitance that a victim sees will increase but not linearly. Figure 55 illustrates this when the radius is $2\mu m$ and the pitch between TSVs is $10\mu m$. The total coupling capacitance was simulated using Synopsys Raphael when different number of aggressors are near a victim TSV. Figure 55 shows that although more TSVs are near the victim, the increase in total coupling capacitance is minor. For example, (d) has two more aggressors than (c), but the

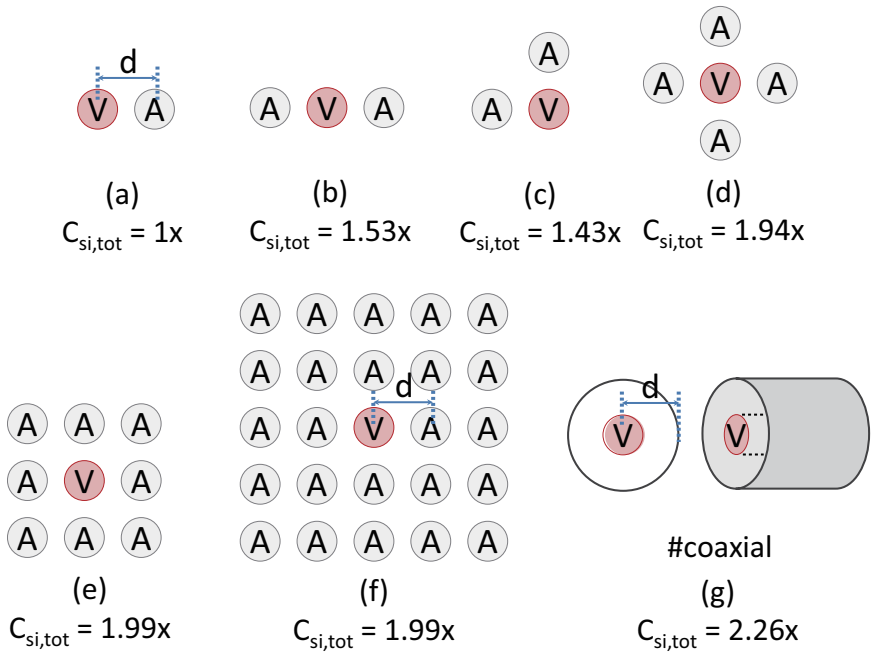


Figure 55: Illustration showing non-linear capacitance increase when the number of aggressors increase, and (g) the maximum limit of coupling capacitance of a TSV.

total capacitance increase is only 0.51x. For (e), four more aggressors are added than (d), but the capacitance increase is only 0.05x. This study shows that Equation 19 cannot be used for multiple TSV coupling analysis. This study also emphasizes that even when there are same number of aggressors, TSV coupling capacitance changes when aggressors are in different locations. For example, Figure 55 (b) and (c) have same number of aggressors but the total capacitance is different by 0.1x. This is because the E-field that forms capacitance changes due to different locations of the TSVs. Thus, note that the coupling capacitance is a function of aggressor locations, as well as a function of distance.

A maximum substrate capacitance limit exists for a TSV victim when the radius (r) and the minimum pitch (P) are given. Even when an infinite number of aggressors are near a victim, the maximum substrate capacitance cannot be larger than that of a coaxial TSV, whose inner conductor radius is r, and the outer conductor, whose inner radius is $P - r$. This formula of a coaxial TSV is shown in in Equation 20 [10].

$$C_{\text{si,max}} = \frac{2\pi\epsilon_0\epsilon_{\text{si}}L}{\ln\left((P-r)/r\right)} \tag{20}$$

Regardless of how many aggressors surround a victim TSV, the total sum of TSV coupling capacitance will be smaller than Equation 20. In other words, no matter how many aggressors surround a victim (as in Figure 55 (f)), the E-field between the victim and the aggressors cannot be formed as strongly as a coaxial TSV (Figure 55 (g)). Although the values of the maximum coupling capacitance will vary on different TSV radii and pitches, when the radius is $2\mu m$ and the minimum pitch between TSVs is $10\mu m$, the maximum capacitance will be around 2.26x. Figure 56 shows how the maximum neighbor capacitance is limited in two different TSV technologies. For a given victim, aggressors are placed at the nearest to a victim in the given pitch, and the number of aggressors are increased. Notice that in these different TSVs, the maximum capacitance rule applies. In summary, the capacitance sum between a victim and the aggressors has a physical limit and cannot be larger than Equation 20.

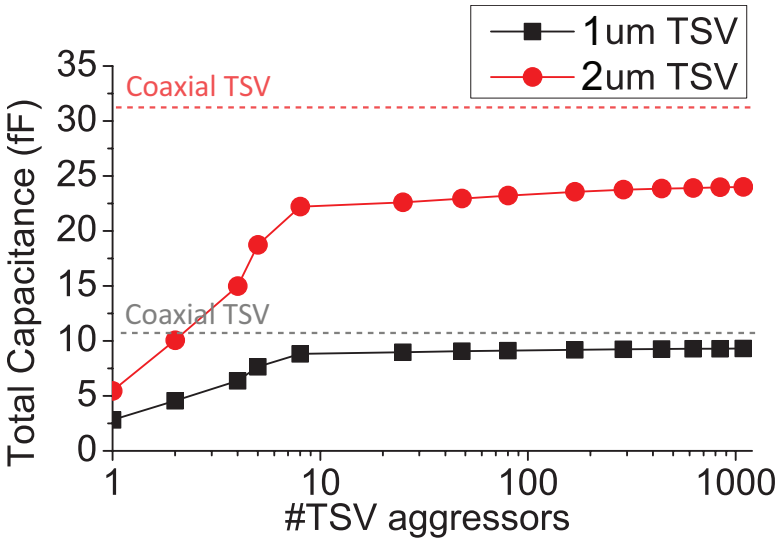


Figure 56: Total capacitance of a victim when # of aggressors increase in two TSV technologies: 1/3/12μm and 2/5/20μm. (radius/pitch/height)

3.3.2 Neighbor Effect on TSV Coupling

Unlike the common belief that only the nearest aggressors impact TSV coupling, TSV coupling occurs even between the non-neighboring aggressors. Assume a simple layout

where a victim TSV is neighboring two aggressor TSVs in a straight line (see Figure 57 (a)). Modeling was performed using the proposed model in Section 3.4.1 and the model was validated using Ansys HFSS. It is intuitively thought that the far aggressor will not affect coupling because a closer neighbor is near by. However, Figure 58 shows that the far aggressor affects as much coupling voltage (139.6mV) as the close aggressor (184.6mV) when 1GHz signal is applied in 45nm transistors. This is because the far aggressor also has non-negligible amount of capacitance between the victim (close aggressor: 9.46fF, far aggressor: 4.14fF, see Figure 59 Case 3). Though the close aggressor shields some of the E-field between the victim and the far aggressor, E-field detours the first aggressor and forms capacitance between the far aggressor and the victim (see Figure 57 (b), field distribution simulated using Ansys Q3D). In addition, despite the far aggressor has less than 50% capacitance of the close aggressor, V_{far} reduces by only 40mV. This is because of the complex coupling network that TSVs compose explained in [69].

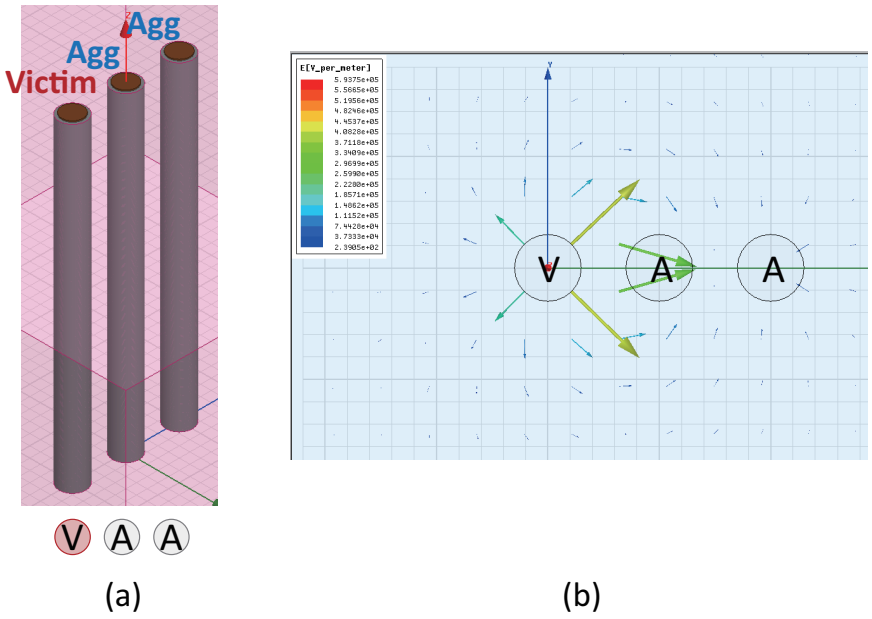


Figure 57: Neighbor Effect. (a) Two aggressor model in HFSS, (b) the E-field distribution between the TSVs.

In addition, neighbor TSVs reduce the capacitance of other TSVs. Figure 59 describes the far aggressor impact on capacitance. Assume there are only two TSVs as Case 1 and

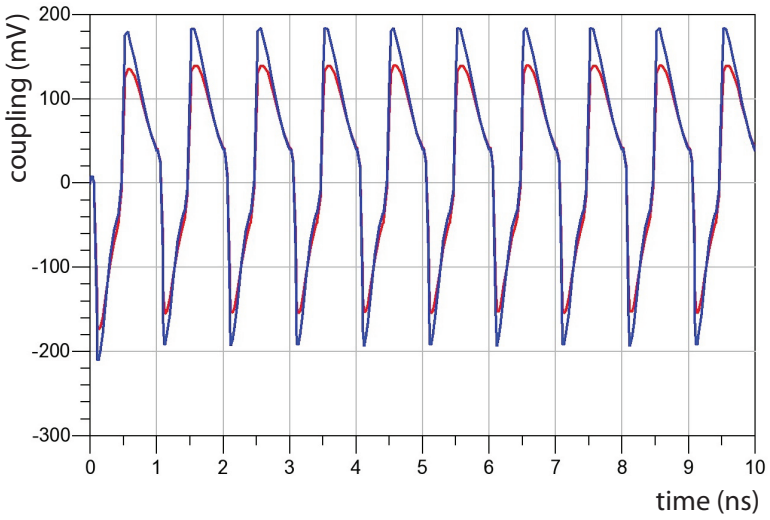


Figure 58: Coupling voltage of the near (blue) and far (red) aggressors shown in Figure 57.

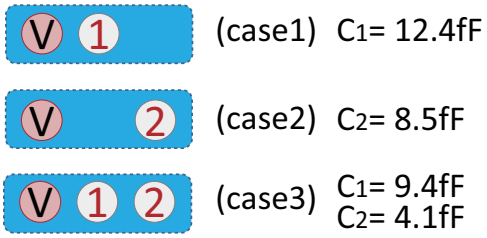


Figure 59: Neighbor Effect case study on how neighbor TSVs affect other aggressors.

Case 2. Each capacitance is 12.4fF (near aggressor) and 8.5fF (far aggressor). However, when two aggressors are together (Case 3), the coupling capacitance of both aggressors decreases to 9.4fF and 4.1fF. This is because the TSVs correlate to each other and create a new E-field distribution. This study will call this the "Neighbor Effect". Using the Neighbor Effect, to reduce the coupling capacitance between an aggressor and a victim, adding another TSV near the original aggressor will help in reducing the capacitance of both the original aggressor and the new TSV. Described in Equation 20, since there is a physical limit to the total coupling capacitance, no matter how many TSV neighbors are added, the total capacitance will be smaller than a certain value. Therefore, it is proven that the coupling capacitance is a function of distance, location, and also a function of neighbors [64, 70].

3.4 Multi-TSV Coupling Extraction

This section proposes a compact multiple TSV-to-TSV coupling model and an extraction algorithm for full-chip analysis.

3.4.1 Compact Multi-TSV Coupling Model

[9] proposed a multiple-TSV model that can be used when performing coupling analysis. However, this model consists of many RLC components even when modeling few TSVs. Thus, this section proposes a compact multi-TSV-to-TSV coupling model that can be easily used on full-chip analysis. Figure 60 (a) shows the original model, and (b) shows the proposed model. Since modern digital systems operate in a clock frequency below 10GHz, the proposed model is targeted to be valid in this range.

3.4.1.1 Silicon Substrate ($C_{\rm si}$ and $R_{\rm si}$) and Model Simplification

To describe the formulas used in the proposed model, the concepts used in [9] are explained first. Assume three aggressors ($N = 3$) are near a victim. An $N + 1$ system considers to become N-conductor transmission line. Using the multi-conductor transmission line theory, a TSV must be assumed as the reference. Thus, this will be assumed to be the victim TSV (#0). Therefore, the victim TSV does not have inductance and only have resistance. A TSV is expressed as a resistor ($R_{\rm TSV}$) and an inductor ($L_{\rm TSV}$) in series. A SiO_2 liner surrounds the TSV for isolation and is expressed as a capacitor ($C_{\rm ox}$). Silicon substrate can be expressed as a resistor ($R_{\rm si,ij}$) and a capacitor ($C_{\rm si,ij}$) in parallel, of which is the resistance and the capacitance between aggressor i and aggressor j. When $i = j$, it is the resistance and the capacitance of the substrate between the victim and the aggressor.

For $R_{\rm si,ii}$ and $C_{\rm si,ii}$, the process starts by calculating $L_{\rm si,ij}$, which is the substrate inductance between two TSVs. $\mathbf{L}_{\rm si}$ is expressed in matrix ($[L_{\rm si}]$), and consists of self-loop inductance and mutual-loop inductance. By definition, $L_{\rm si,ii}$ indicates the substrate inductance between the victim and the aggressor i. The following equations describe how to

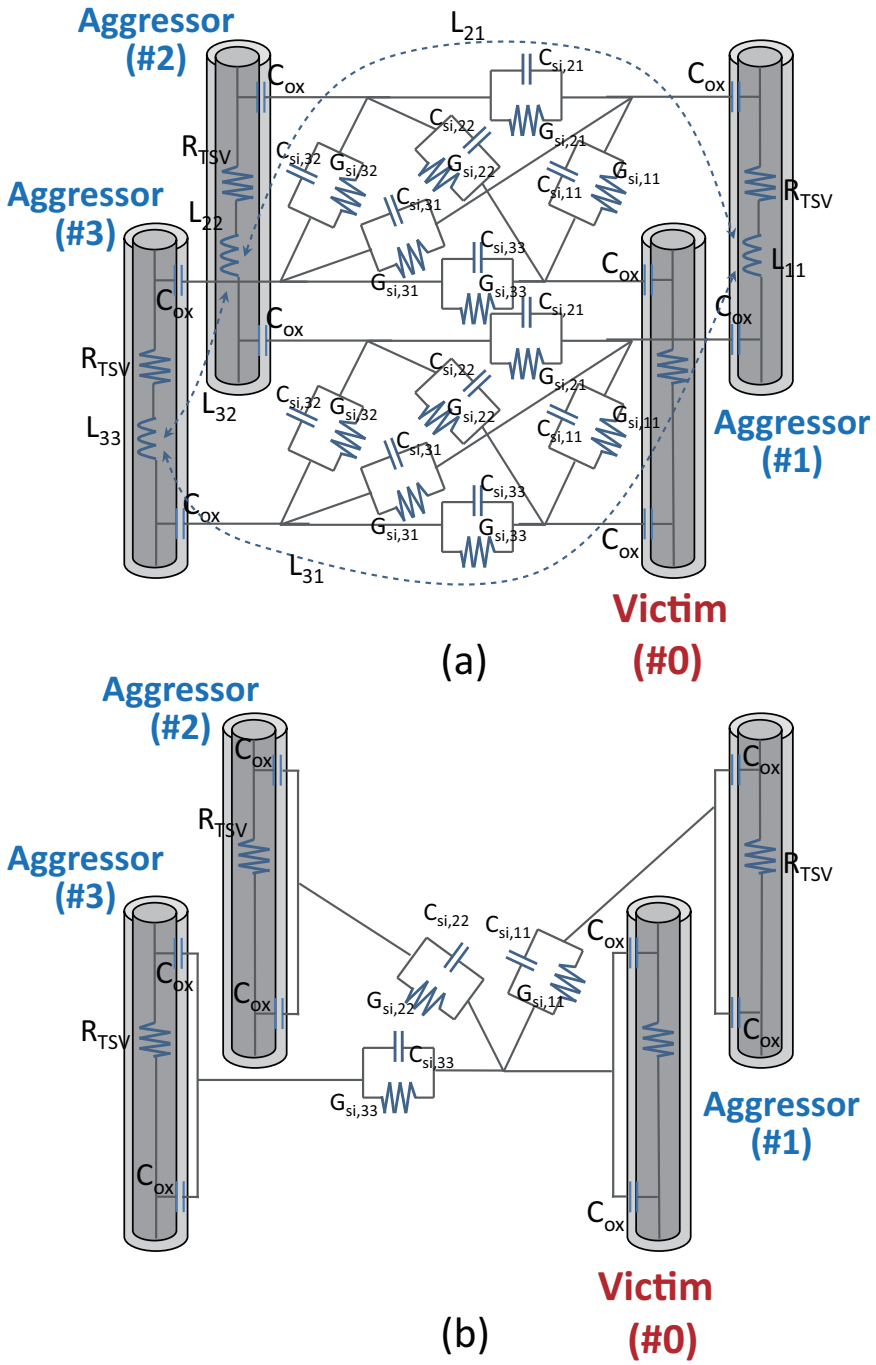


Figure 60: (a) Original model proposed in [9], and (b) the proposed compact TSV model for full-chip analysis.

calculate these values,

$$L_{\mathrm{si,ii}} = \frac{\mu L}{\pi} \ln \left[\frac{P_{\mathrm{i0}}}{r + t_{\mathrm{ox}}} \right] \tag{21}$$

$$L_{\text{si,ij}} = \frac{\mu L}{2\pi} \ln \left[\frac{P_{\text{i0}} P_{\text{j0}}}{P_{\text{ij}}(r + t_{\text{ox}})} \right] \tag{22}$$

where P_{i0} is the pitch between the victim TSV ($\#0$) and the aggressor TSV($\#i$), and P_{ij} is the pitch between two aggressor TSVs ($\#i$, and $\#j$). By the relation between the inductance matrix and the capacitance matrix in a homogeneous medium [61], matrix $\mathbf{C}_{\text{si}}$ is calculated,

$$\mathbf{C}_{\text{si}} = \mu_0 \epsilon_0 \epsilon_{\text{si}} L^2 {\mathbf{L}_{\text{si}}}^{-1} \tag{23}$$

where $\mathbf{C}_{\text{si}}$ and its inner components $C_{\text{si,ij}}$ are defined as Equation 24

$$[C_{\text{si,ij}}] = \begin{bmatrix} \sum_{k=1}^{N} C_{1k} & -C_{12} & \dots & -C_{1N} \\ -C_{21} & \sum_{k=1}^{N} C_{2k} & \dots & -C_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{N1} & C_{N2} & \dots & \sum_{k=1}^{N} C_{Nk} \end{bmatrix} \tag{24}$$

and the conductance matrix $\mathbf{G}_{\text{si}}$ is defined as Equation 25.

$$\mathbf{G}_{\text{si}} = \frac{\sigma}{\epsilon_0 \epsilon_{\text{si}}} \mathbf{C}_{\text{si}} \tag{25}$$

In the proposed model, only $C_{\text{si,ii}}$ and $G_{\text{si,ii}}$ ($R = 1/G$) are used. The other RLC components are reduced. This is reasonable because the impact between a victim and an aggressor are considered and not the impact between two different aggressors. Using the proposed model, RLC count is reduced by 60% when N=3. The RLC count reduces more as N increases. Despite the RLC reduction, the proposed model is shown accurate described in Section 3.4.1.5.

3.4.1.2 Inductance Modeling (L_{ij})

Self inductance and mutual inductance is removed in the proposed TSV model. However, this is reasonable due to the following reasons: First, the TSV inductance, which is in few

tens of pH range, have negligible impact on delay and coupling noise on the frequency range of digital circuits ($<$ 10GHz). For example, the impact of TSV inductance (self and mutual) on delay and coupling noise is less than 2% in 1GHz clock. This means that in digital circuits, capacitive coupling is the dominant coupling factor and inductive coupling is almost negligible. This is shown in Figure 61 that even though inductance is removed in the proposed model, S-parameter comparison shows a good correlation between the 3D-EM simulator model and the simplified model. For inductance to impact on delay and coupling, it requires to be in the range of nH in the frequency target. However, for example, 1nH is an inductance that can be seen in a wire that is longer than 1mm. Despite that TSV scaling leads to possibilities of TSV inductance increase due to pitch decrease, note that the TSV size also scales as TSV pitch reduces. Thus, TSV inductance remains in the pH range despite the technology scaling. Due to these reasons, since inductive coupling is almost negligible, inductance is removed from the proposed model.

3.4.1.3 Resistance of the TSV (R_{TSV})

In TSVs, skin effect occurs on the AC current that flows inside. Thus, as frequency increases, R_{TSV} starts increasing from a certain frequency point. Equation 26 describes the formula for R_{TSV}

$$R_{\mathrm{TSV}} = \frac{L}{2\pi r}\sqrt{\frac{\pi f \mu_0}{\sigma_c}} \tag{26}$$

where μ_0 denotes the permeability of free space, f the frequency, and σ_c the conductivity of copper, respectively. For example, in a 5μm diameter TSV, the resistance starts increasing from 700MHz due to skin effect. As TSV diameter scales, the frequency that starts increasing R_{TSV} due to skin effect will increase. This is because smaller TSVs (in diameter) will approach the skin depth in a higher frequency than in larger TSVs.

3.4.1.4 Capacitance of the liner (C_{ox})

For C_{ox}, SiO_2 liner surrounding the TSV can be modeled as capacitance of the liner itself and the MOS capacitance [30] of the TSV in parallel

$$C_{\text{ox}} = \frac{C_{dep} C_{liner}}{C_{dep} + C_{liner}} \tag{27}$$

$$C_{\text{liner}} = \frac{2\pi\epsilon_0\epsilon_{ox} L}{\ln\left(\frac{r+t_{ox}}{r}\right)} \tag{28}$$

$$C_{\text{dep}} = \frac{2\pi\epsilon_0\epsilon_{si} L}{\ln\left(\frac{r+t_{ox}+t_{dep}}{r+t_{ox}}\right)} \tag{29}$$

where t_{dep} is the thickness of the depletion region. In the assumption, when substrate doping is $10^{15}/cm^3$, note that a depletion region always exist around TSVs in digital systems that operate between 0V and VDD.

3.4.1.5 Model Validation

The proposed model is validated by first placing aggressor TSVs around the victim TSV randomly in a fixed space. Then, modeling is performed using 3D EM solver HFSS, and also a SPICE netlist is generated based on the proposed compact model. HFSS provides accurate models in cost of significant runtime, E.g., generating a 10 TSV model in HFSS takes more than one hour, while the proposed SPICE model generation takes less than a second. Therefore, HFSS modeling is not feasible for full-chip analysis, and SPICE is a good approach to handle many TSVs in full-chip. 10 layouts are generated for each sample cases, and then the S-parameter of these two components are compared and the maximum error of insertion loss is reported. Figure 61 shows the S-parameter comparison when N=3, and Table 5 shows the validation result. It is shown that the proposed model is very accurate, even in a multiple TSV structure, by reporting the maximum difference in insertion loss less than 0.02dB. This chapter do not considers the impact of inter-tier TSV-to-TSV coupling. This is because many metal interconnects are placed between the

inter-tier TSVs, and this shields the E-field between TSVs on the different layers to form capacitance.

Table 5: Model validation on general layouts

TSV dimensions (μm)			# TSVs	Average err (dB)	Max. err (dB)
Radius	Pitch	Height			
2	5	30	6	0.008	0.016
			8	0.011	0.015
			10	0.008	0.014
			12	0.011	0.015
		60	6	0.009	0.015
			8	0.011	0.016
			10	0.011	0.015
			12	0.008	0.014
4	10	30	6	0.010	0.016
			8	0.009	0.014
			10	0.011	0.017
			12	0.011	0.018
		60	6	0.010	0.017
			8	0.009	0.014
			10	0.010	0.015
			12	0.008	0.014

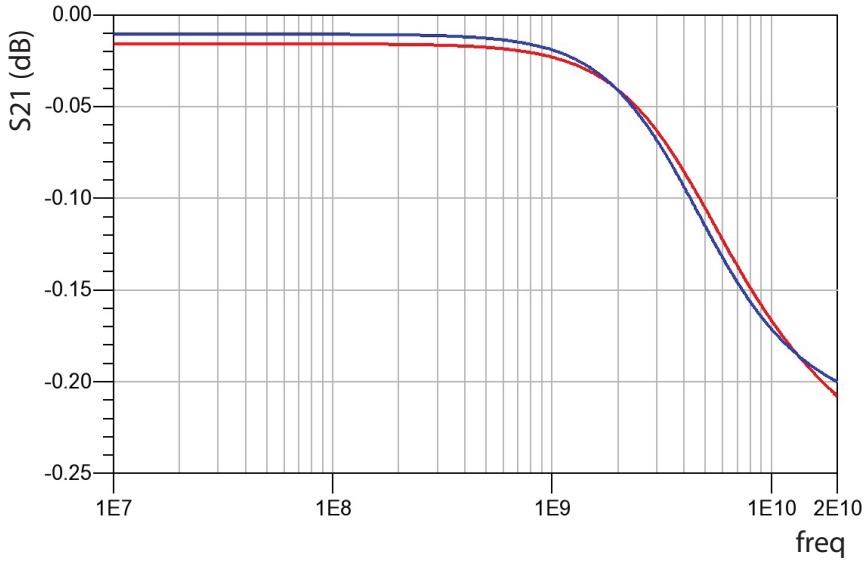


Figure 61: S-parameter comparison between the proposed model and HFSS (red: HFSS, blue: proposed model)

3.4.2 Extraction Algorithm

In the previous discussions (Section 3.3.1 and 3.3.2), it was shown that TSV coupling capacitance is a function of distance, location, and neighbor aggressors. To extract TSV-to-TSV coupling capacitance accurately, an approach considering only the closest neighbor or limiting the maximum target distance to calculate coupling capacitance cannot be used. Therefore, this study proposes an algorithm that considers distance, direction, and Neighbor Effect all in a holistic manner when extracting the coupling capacitance for all nets in the layout for full-chip analysis. Algorithm 2 describes this.

Algorithm 2: Multiple TSV-to-TSV capacitance extraction

```
1  Algorithm: Multiple TSV-to-TSV capacitance extraction
2  Locate all TSVs by its coordinate (x,y);
3  while for all victim TSVs do
4  |  For all neighbor TSVs, calculate the Euclidean distance and sort by the closest
   |  distance to the victim;
5  |  Choose N aggressors that is closest to the victim;
6  |  Calculate the coupling capacitance of the
   |  N aggressors using the formula in Section 3.4.1;
7  |  if The calculated TSV capacitance is higher than C then
8  |  |  Generate a coupling network between the aggressor and the victim;
9  |  else
10 |  |  Assign the TSV coupling capacitance to be zero;
11 |  end
12 end
```

In an actual layout, any TSV can become a victim from noise. Therefore, the proposed full-chip 3D SI analysis flow described in Section 3.5.1 analyzes the coupling noise in every net of the chip. Thus, the proposed algorithm must be performed for every TSV. From a given layout, the (x,y) coordinate of each TSV is first extracted. Starting from the very first TSV of the layout, this is assumed to be a victim and all neighbor aggressor TSVs are sorted by the closest Euclidean distance to the victim. Then, N neighbor aggressor TSVs (N: a significantly large number) are chosen from the sorted result that are closest from the victim and the capacitance between the victim and the chosen aggressors are calculated.

Once the capacitance of the aggressors are calculated, a coupling network is generated between the victim and the aggressor that the capacitance is higher than a certain value (e.g., C $>$ 0.01fF).

A significant number of aggressors (N: more than 100) are chosen after sorting to guarantee that any aggressors that are physically far but meaningful (far from the victim but does not have any closer neighbors between the victim) are not neglected. Figure 62 illustrates this idea. Unless a certain number of aggressors are chosen for analysis, it can accidentally miss the valid aggressors that must be considered for extraction. For example, when N=10, the aggressor circled in blue is ignored. This can be considered only when N is bigger than 114. Therefore, N must be a big number that can consider all effective neighbors in the layout. This step is repeated for every TSV in the layout, and the corresponding coupling network is created for each victim TSV.

The proposed algorithm considers all aggressors that affect the victim. Using the algorithm with the proposed TSV model, the Neighbor Effect is successfully considered. In a layout, it is not only the distance, but also the location and the neighbors that is important. Since the proposed algorithm calculates the coupling capacitance from a very large number of aggressors, not by distance, it does not neglect any aggressors that must be considered.

3.5 Full-chip Analysis

Using the proposed extraction flow, this section performs full-chip SI analysis and compare the results to [45].

3.5.1 Full Chip 3D SI Analysis Flow

Since existing SI analysis tools cannot analyze 3D circuits accurately, this research modified the 3D SI analysis flow in [45] to implement results. First, the SPEF file is extracted for each die using RC extraction tool. Then, scripts that implements the algorithm developed in Section 3.4.2 are executed to create the SPEF file of TSV parasitics. Then, a top-level verilog file is created. Once these files are prepared, Synopsys PrimeTime is used to read

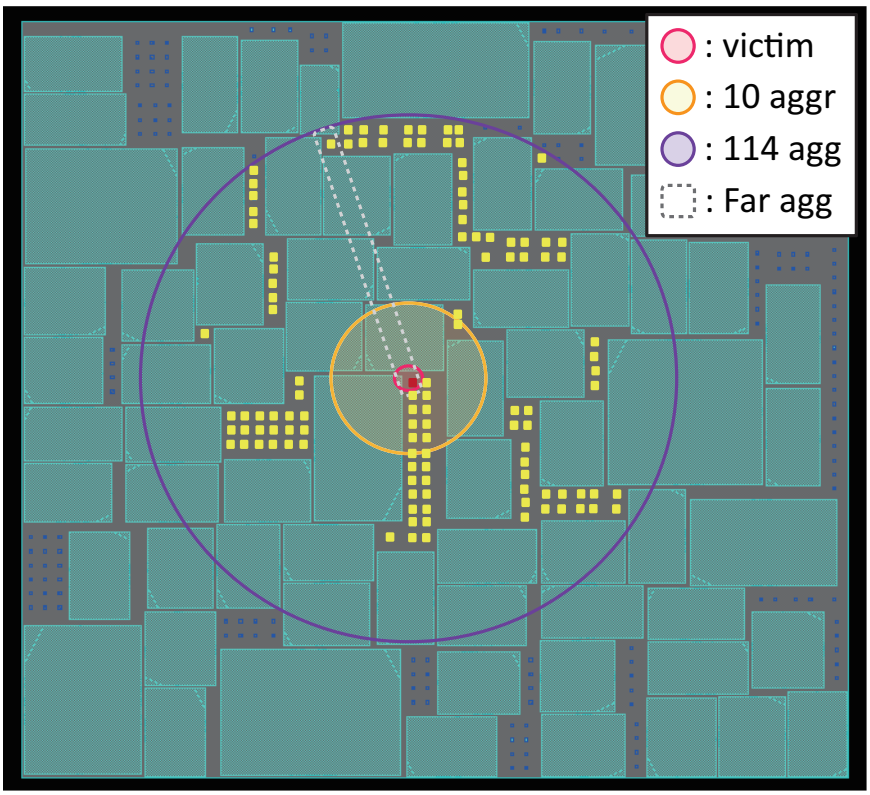


Figure 62: Comparison between a small N (10 aggressors) and a large N (114 aggressors) in the proposed algorithm.

the verilog file, and a top-level stitched SPEF file is createed that contains RC information of all dies and the TSVs. This step inserts the extracted coupling network from Section 3.4.2 into SPEF file. Then, the stitched SPEF file and generate a SPICE netlist is analyzed for each individual net for performing coupling noise simulation. The SPICE netlist has all the coupling information including wire-coupling, TSV coupling network by the extraction algorithm, and the aggressor signal and the victim driver models. HSPICE is executed on each net one by one, assuming the aggressors are switching and report the peak noise at each port.

3.5.2 Design and Analysis Results

FFT 256-8 is used as the benchmark, which is a 256 point with 8 bit precision, real and imaginary FFT. The circuit has 140K gates and 211 TSVs. The design is a 2-tier 3D IC based on Nangate 45nm technology. The designs were based on the Cadence Encounter

design flow to generate 3D layouts [42]. In Figure 63 and Table 6, coupling analysis results of top-hierarchy nets, which are around 3K, are shown and compared with [45]. E.g., in Figure 63, w/o coupling analysis shows around 800 nets in the 0-100mV bin, and [45] and results in this study show around 700 nets in the 0-100mV bin.

Based on the results, the following impacts can be observed: First, for coupling noise, both approaches calculate higher coupling noise than w/o TSV coupling (590V). Total coupling noise is the sum of coupling noise voltage that is occurred on each net. Note that when noise voltage occurring on a particular net exceeds a certain threshold, the logic value will be inverted leading to erroneous behaviors inside circuits. More total coupling noise in a layout means that the particular design is more prone to logic failures statistically. Despite [45] is overestimating the coupling capacitance by linear superposition, results in this study shows higher total noise voltage. The total coupling noise is 732V using the flow in [45] and 787V in results of the proposed flow. This is because the proposed model considers more neighbor aggressors than [45] that should not be ignored. Note that 196.65V (787.42V - 590.77V) is the noise that has been generated due to TSV coupling. In this chapter, this TSV-induced noise will be defined as "3D noise".

Table 6: TSV coupling impact on crosstalk and timing. Coupling noise in (V), longest path delay in (ns), and total negative slack in (ns)

	W/O coupling	W/ coupling [45]	W/ coupling (this study)
Footprint (mm^2)	0.7954	0.7954	0.7954
Tot. coupling noise	590.77 V	732.75 V	787.42 V
Longest path delay	2.734 ns	3.165 ns	2.852 ns
Total negative slack	-61.65 ns	-115.07 ns	-75.24 ns

Second, for timing analysis, because [45] overestimates the total coupling capacitance, it also overestimates the timing degradation by TSVs as well. The proposed method saves a significant timing margin by using an accurate TSV model. Note that the longest path delay (LPD) and total negative slack (TNS) depends on the total capacitance formed between aggressor TSVs, and coupling noise depends on the number of aggressors formed between

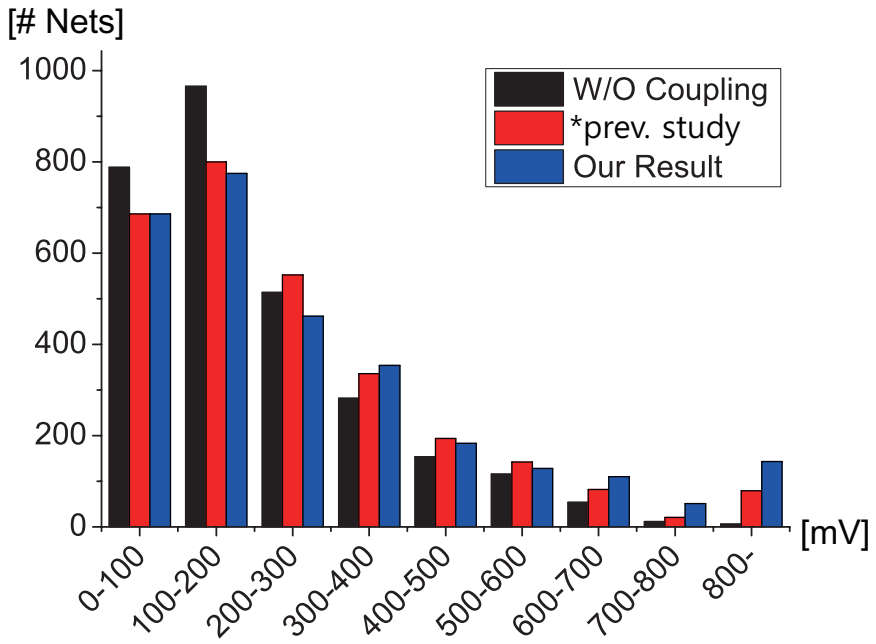


Figure 63: Coupling analysis result. X axis denotes the noise voltage bins, and Y axis denotes the number of nets contained in the specific bin. Previous study refers to [45]

the victim. TNS is the sum of the negative slack for all paths that fail any timing constraint.LPD tells the designer what the maximum clock period could be, and TNS shows how far off the circuit is from reaching timing closure. Figure 64 shows how noise and delay trend is different compared to [45]. In terms of timing, the most important factor is the total capacitance. Despite [45] considers less aggressors, it overestimates capacitance. Thus, the total capacitance formed from aggressors are larger than that of the analysis proposed (18fF $>$ 11fF). However, in terms of noise, the most important factor is the number of effective aggressors. Note that a small capacitance formed between the aggressor and victim could lead to a big coupling voltage (Section 3.3.2). Since the proposed analysis considers more effective aggressors, it analyzes more coupling noise than in [45].

3.6 Impact of Process Parameters on TSV Coupling

This section studies the impact of process parameters on TSV coupling in terms of coupling coefficient and full-chip impact. For the full-chip results, the TSV parameters on the design performed on Sec. 3.5 are varied to gain understanding of how these parameters impact the

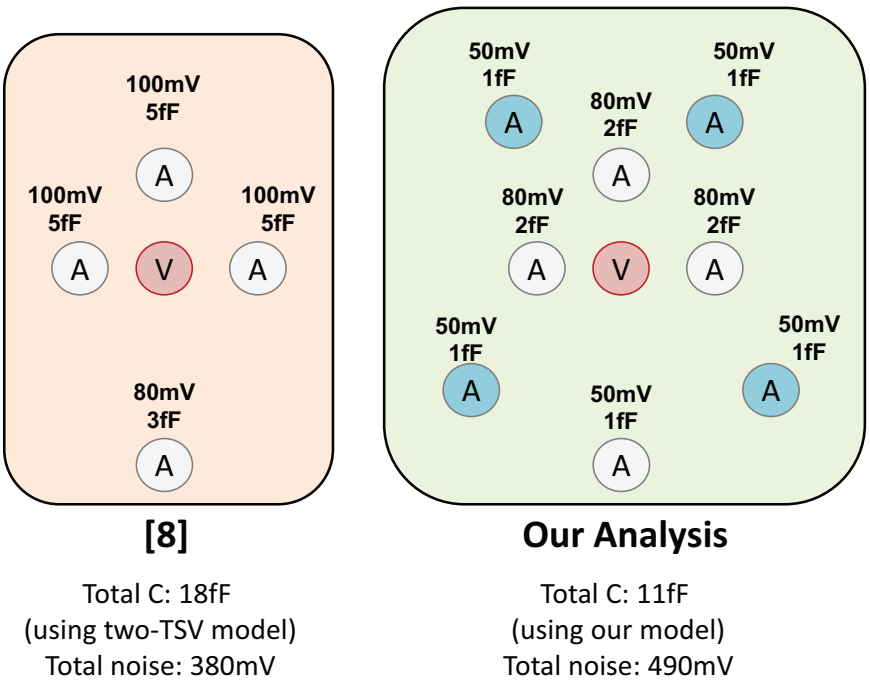


Figure 64: Why delay and noise trend is different. Left shows the analysis in [45], and right shows analysis of this work.

full-chip design.

3.6.1 TSV Height

The impact of TSV height is firstly studied. TSV height is determined by the die thickness. With a shorter TSV height, the TSV resistance and capacitance reduces, which is good for reducing TSV induced coupling. Therefore, die thinning is one of the keys to a good TSV technology. Here, this section analyzes when TSV height is from $20\mu m$ to $100\mu m$. It is seen that the coupling coefficient increases monotonically with the TSV height as expected (Figure 65). This is because all TSV parasitics are linearly proportional to TSV height. In terms of full-chip results (Table 7), TSV height increase leads to additional 3D noise. Notice that the 5x TSV height increase does not lead to 5x coupling noise increase due to the complicated TSV coupling network [69]. Comparing $20\mu m$ and $100\mu m$ TSVs, 27.1% 3D noise increase is seen due to TSV height increase.

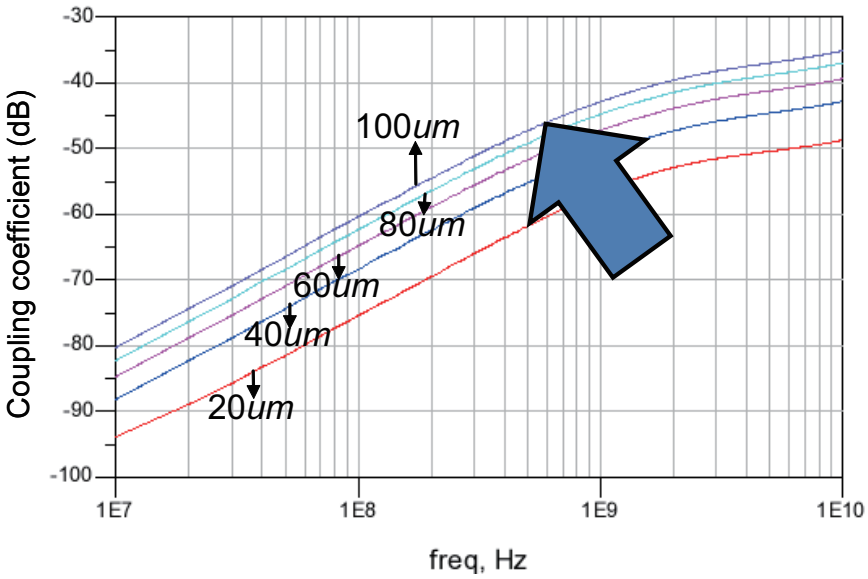


Figure 65: S-parameter simulation of coupling coefficient with different TSV heights (20-100μm).

Table 7: Full-chip 3D noise: Impact of TSV parameters.

TSV height	20μm	40μm	60μm	80μm	100μm
3D noise (V)	155.1	169.2	180.3	189.5	197.1
Ratio (%)	0	9.0	16.2	22.1	27.1
Liner thickness	0.1μm	0.2μm	0.3μm	0.4μm	0.5μm
3D noise (V)	204.2	194.9	188.6	184.0	180.3
Ratio (%)	0	-4.6	-7.6	-9.9	-11.7
TSV diameter	2μm	4μm	6μm	8μm	10μm
3D noise (V)	180.3	199.6	226.0	251.2	256.1
Ratio (%)	0	10.7	25.3	39.3	42.0

3.6.2 Liner Thickness

TSV liner also has a significant impact on TSV capacitance. Thickness of TSV liner varies from 0.1μm to 0.5μm and the coupling coefficient is reported. In Figure 66, as the liner thickness is increased, the coupling coefficient decreases in the low frequency region but not in the high frequency region. Liner capacitance contributes only in the low frequency region due to its size and geometry inside the coupling network. Thus, coupling impact due to liner capacitance will reduce as the operating frequency increases. In this full-chip study (Table 7), changing the liner thickness from 0.1μm to 0.5μm leads to -11.7% 3D coupling

noise difference.

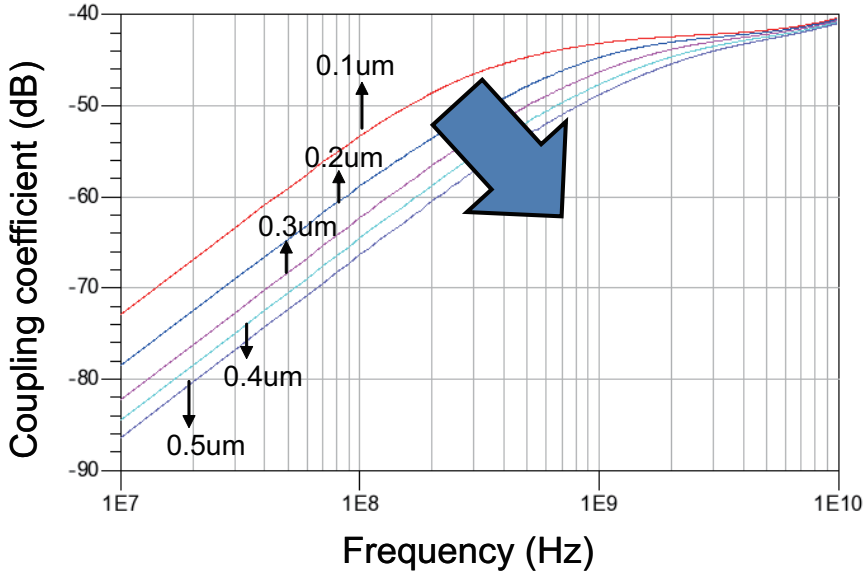


Figure 66: S-parameter simulation of coupling coefficient with different liner thickness (0.1-0.5μm).

3.6.3 TSV Diameter

TSV diameter affects both the TSV capacitance and the resistance. A bigger TSV diameter helps to reduce the TSV resistance. However, due to the increased TSV oxide area, the TSV capacitance will increase significantly. Usually the TSV resistance is very small (50mΩ). Thus, TSV capacitance ($50fF$) is usually the dominant factor of the TSV parasitics. Since the TSV capacitance has a dominant role in the TSV coupling, it is expected that a bigger diameter will increase the coupling noise. Figure 67 shows the analysis results. It is shown that with bigger TSV diameter, coupling coefficient increases as expected. In full-chip results (Table 7), TSV radius change showed the highest noise difference (42%) within the given range of variation in this study. In addition to the TSV capacitance increase when TSV diameter increases, note that TSV-to-TSV distance also reduces, which further enhances the TSV-to-TSV coupling.

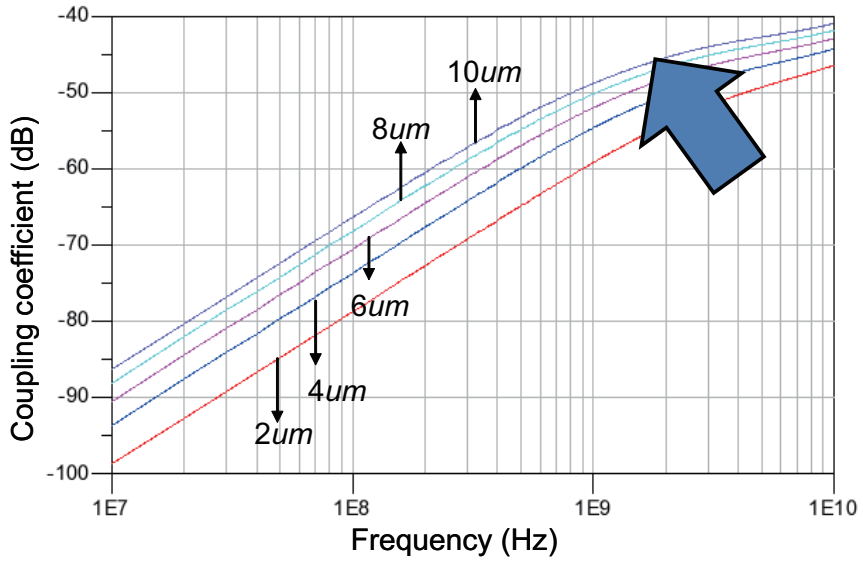


Figure 67: S-parameter simulation of coupling coefficient with different TSV diameters (2-10μm).

3.7 Impact of Process Parameters on Delay

This section studies the impact of TSV process parameters on timing and delay. To analyze TSV impact on delay, this section proposes an "Impedance Load Analysis" method.

3.7.1 Analysis Structure for Single Net Delay Study

Figure68 shows the test structure for the single net delay study on 3D TSV. In this model, Driver (std. cell) #1 from the left bottom drives the victim TSV, and the delay at the node on Receiver #1 is measured. A neighbor TSV and its driver and receiver on its right is also included to see the impact of neighbor TSVs on delay. Note that since this is a delay study, Driver #2 is not switching. Driver #1 size varies from the minimum (1x) to the biggest (16x), and the receiver size becomes the same as the driver size.

3.7.2 Impact of TSV Height, Liner Thickness, and TSV Radius

Figure 69 shows the delay impact of TSV height, liner thickness, and TSV radius, respectively. As in Sec. 3.6, a similar trend in delay is shown as well. When TSV height increases, both TSV resistance and TSV capacitance increases. Thus, delay increases as TSV height

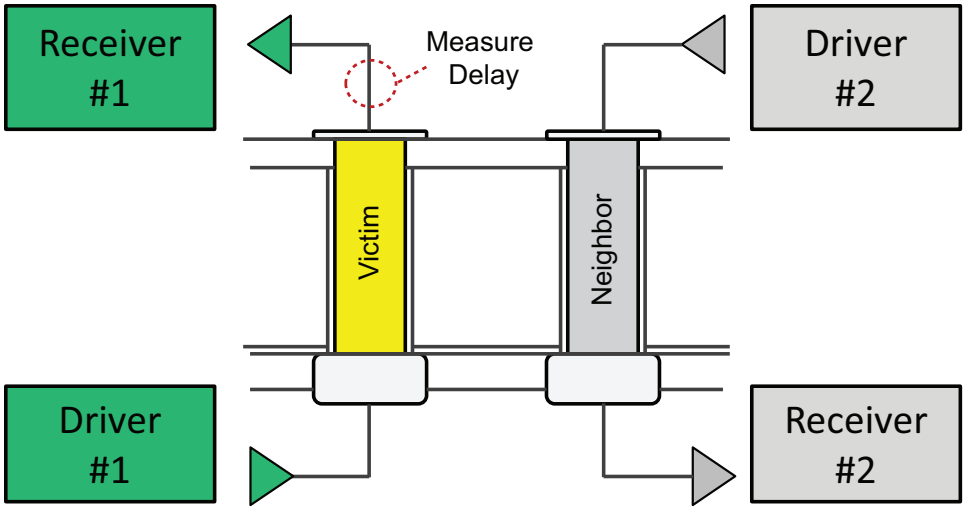


Figure 68: Single net delay analysis model of a TSV having one neighbor TSV.

increases as in (a). When the SiO_2 liner thickness increases, TSV resistance remains the same but TSV capacitance reduces. Therefore, delay decreases as liner thickness increases as in (b). When TSV radius increases, despite that TSV resistance reduces, TSV capacitance increases significantly. Thus, delay increases as TSV radius increases as in (c). Note that drivers stronger than 8x will not see a significant delay impact from TSV parameter change. In other words, drivers must be strong enough to minimize the delay impact on 3D TSV nets due to the significant capacitance load that a driver sees.

3.7.3 Impact of TSV Pitch

This section performs the same single-net experiment changing the pitch between TSVs from 10μm to 50μm and shows the results in Figure 70 (a). It is shown that changing TSV pitch does not impact much on reducing the delay of the victim receiver (-7.9% reduction in 1x driver). However, this is different when a 2D net is distanced from an aggressor. To compare the impact of delay reduction in 2D and 3D due to neighbor pitch change, an experiment where a 2D wire has the same dimension as a TSV is performed, in which the permitivity of the dielectric is the same as the silicon substrate. Figure 70 (b) shows the 3D vs. 2D delay comparison. Both 1x driver size is used in both experiments. In this experiment, TSV delay will be defined as 3D delay, and 2D wire delay that has the same

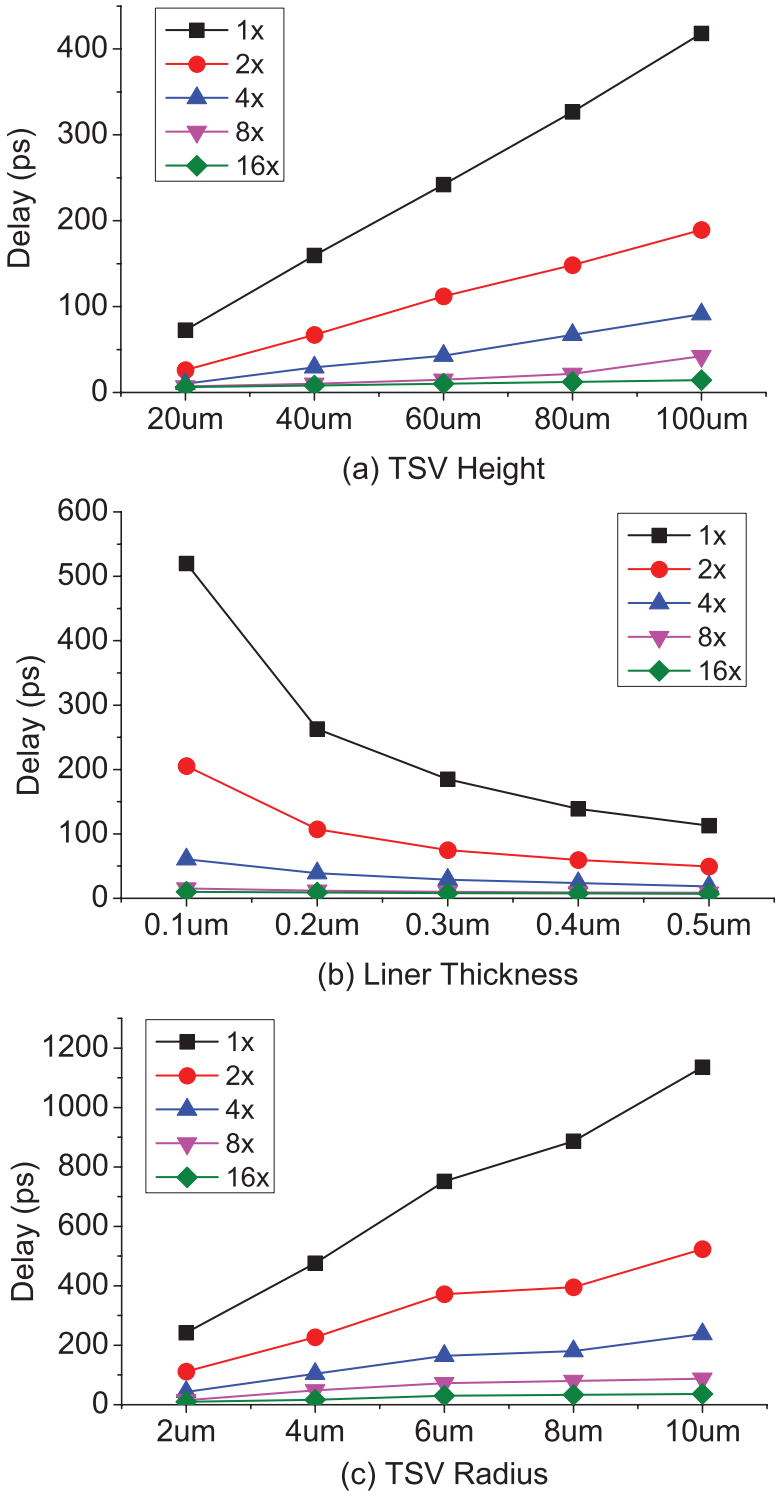


Figure 69: Delay impact on various TSV parameter change when driver (std. cell) size changes (1x – 16x): (a) TSV height, (b) Liner thickness, and (c) TSV radius.

dimensions as TSV will be defined as 2D delay, respectively. Here, two important findings are reported: (1) At the same pitch, 3D delay is always higher than 2D delay. In 10μm neighbor pitch, the 3D delay is almost 2x of 2D delay (242ps $>$ 110ps), and this difference increases as the pitch increase. In 50μm pitch, this delay difference is more than 4x (225ps $>$ 54ps). (2) Unlike in 2D, 3D delay does not significantly reduce from increasing the pitch. When the neighbor pitch increases from 10μm to 50μm, 3D delay reduces by only 7.9%, but 2D delay reduces more than 50%. This means that 3D delay is not sensitive to neighbor TSVs unlike in 2D. Note that a similar trend is seen as this in various 2D wire and 3D TSV sizes: 3D delay is always bigger than 2D delay in the same size and less sensitive to distance change. The reason to this discussed in the next section (Sec. 3.7.4.3).

3.7.4 The "Impedance Load" Analysis for Delay Estimation

Calculating the RC delay is a good approach when the delay of a net in normal 2D systems [79] is estimated. When calculating the delay, a net is composed of the resistance of the path and various capacitive loads. These loads are the capacitance load of the receiver, capacitance formed to the GND, and coupling capacitance between paths as in Figure 71 (a). When excluding the resistance for the path, it can be thought that these capacitive loads are the total load that a driver sees in a net as in Equation 30 for delay estimation.

$$Load_{\mathrm{Driver}} = C_{\mathrm{receiver}} + C_{\mathrm{GND}} + C_{\mathrm{coup}} \tag{30}$$

The "Capacitive Load" concept is applicable in normal load conditions where the coupling neighbors are perfectly isolated by a dielectric that its conductivity is almost negligible. However, this cannot be applied to a 3D net with TSVs. In a 3D net, silicon substrate lies between neighbor TSVs that its conductivity is non-negligible. Because of this, silicon substrate introduces an impedance path that is modeled as a resistance ($\mathrm{R_{si}}$, see Figure 71 (b)). Therefore, this study proposes a method of analyzing the delay of a 3D path called the "Impedance Load".

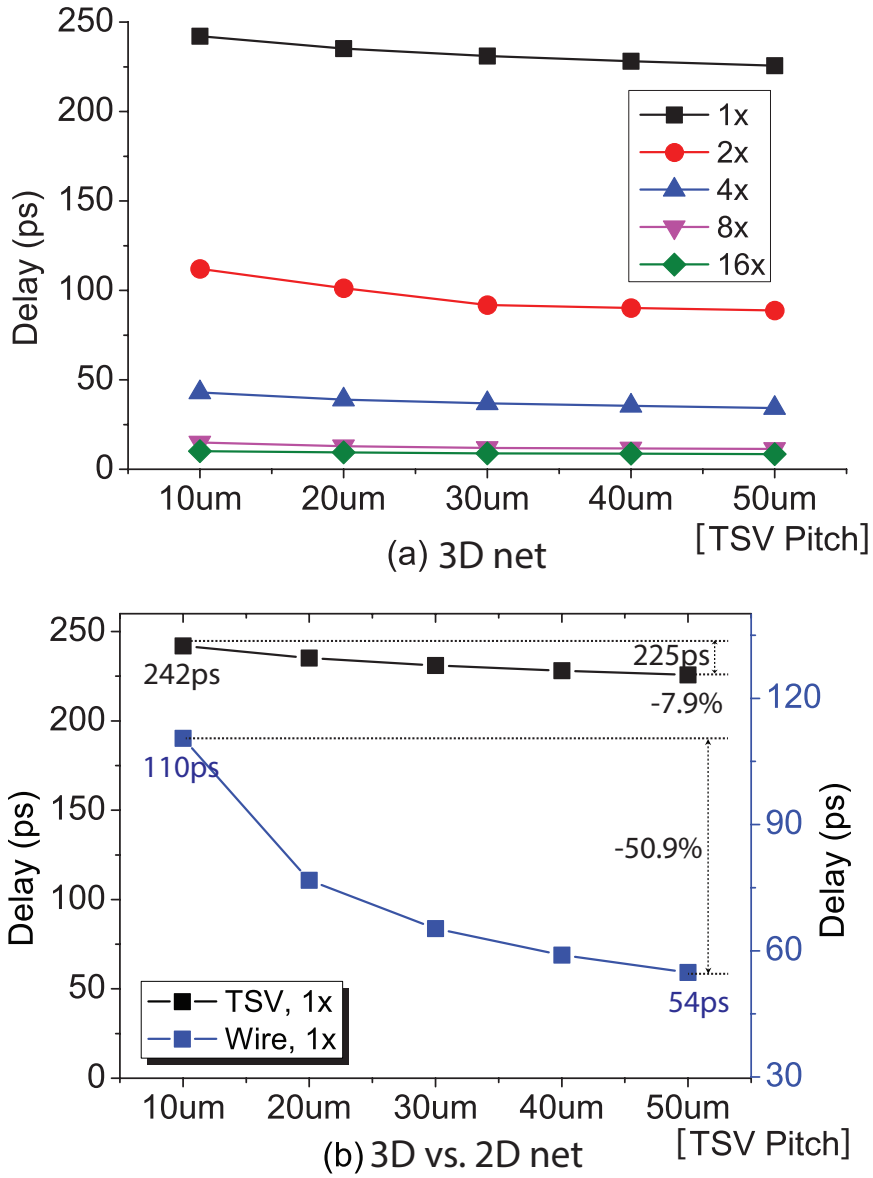


Figure 70: Delay impact when TSV pitch changes: (a) Driver sizes from 1x to 16x, and (b) Comparison between 3D (black) and 2D (blue) when having same dimensions

Using the Impedance Load, Equation 30 changes to Equation 31,

$$Z_{\text{Load}_{\text{Driver1}}} = Z_{\text{receiver}} + Z_{\text{GND}} + Z_{\text{coup}} \tag{31}$$

where all capacitance load transforms into impedance loads. When the loads are expressed as capacitances, the impact of R_{si} cannot be analyzed, but this study can using the Impedance Load analysis method.

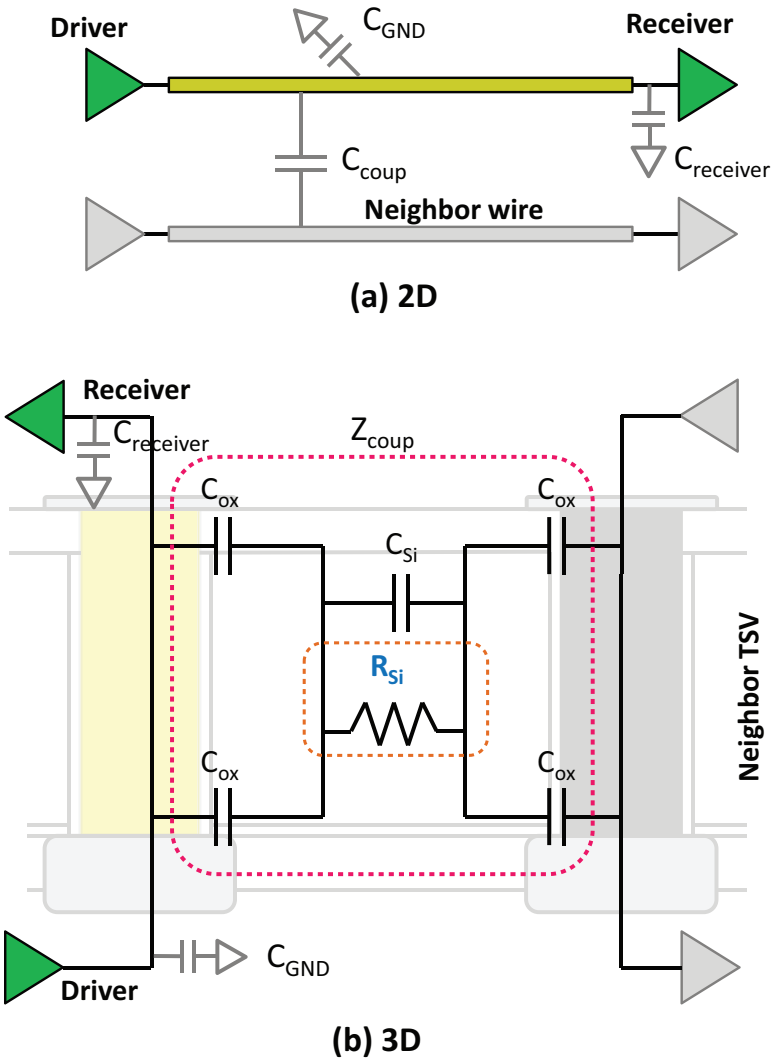


Figure 71: All loads (GND, receiver, and coupling) in (a) 2D net and (b) 3D TSV net.

3.7.4.1 Understanding the Impedance Load for Less Delay

A 2D net example in Figure 72 is described to understand how the Impedance Load concept is used. As in Figure 72 (a), it is seen that the driver sees a load of $\mathrm{C_{load}}$. In the Impedance Load analysis, this capacitive load becomes an impedance load $\mathrm{Z_{C_{load}}}$ ($Z = 1/sC$) as in Figure 72 (b). For example, in 1GHz,

- $\mathrm{C_{load1}} = 1/2\pi$ fF becomes $\mathrm{Z_{C_{load1}}} = 1\mathrm{M}\Omega$
- $\mathrm{C_{load2}} = 10/2\pi$ fF becomes $\mathrm{Z_{C_{load2}}} = 0.1\mathrm{M}\Omega$.

Under the impedance load analysis, a 10x bigger capacitance load ($C_{load2} = 10 \times C_{load1}$) becomes a 0.1x smaller impedance load (($Z_{C_{load2}} = 0.1 \times Z_{C_{load1}}$)) in magnitude. To translate this into a physical meaning, note that a smaller 'Z_{load}' derives more delay. In the perspective of a driver, the voltage swing is a function of the current driving and the load that a driver is seeing ($\Delta V = \Delta I Z$). This means that it requires more current to drive (=change the voltage) a smaller impedance (= higher capacitance) load than a higher impedance (= less capacitance) load. However, since a driver has a limited amount of current driving capability (ΔI), it will take more time to drive a small impedance load than a high impedance load, which small impedance load suffers from more delay.

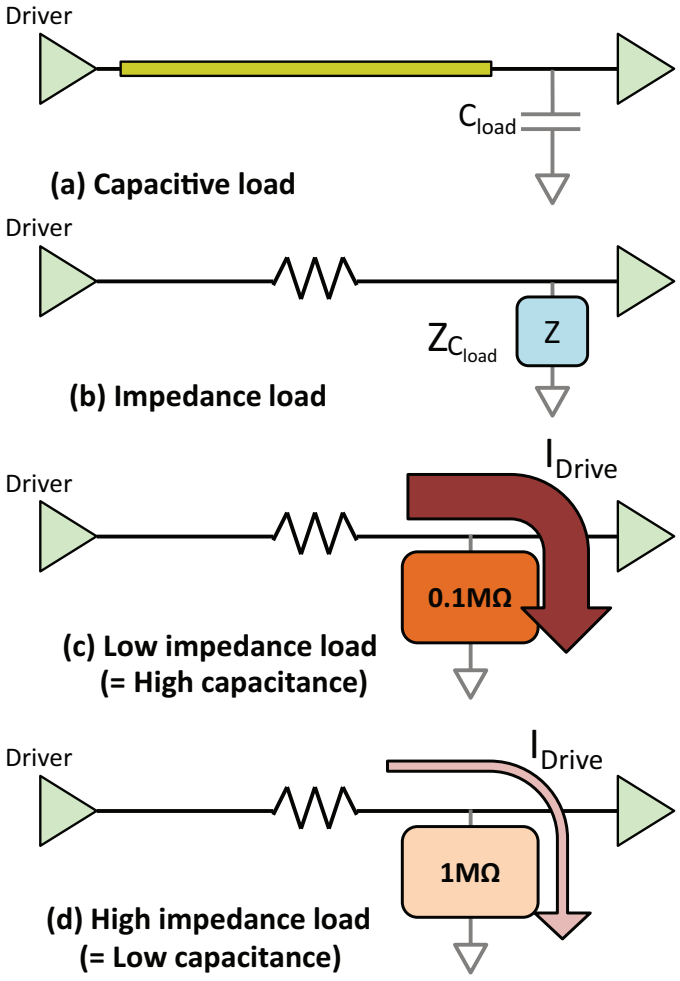


Figure 72: The "Impedance Load" concept. A capacitive load (a), translates to an impedance load (b). Low-impedance load (c) suffers from more delay than high-impedance load (d).

3.7.4.2 *Impedance Load Analysis of a Timing Path*

Using the Impedance Load, the equations are derived for 2D and 3D coupling loads. As shown in Equation 30, a coupling load C_{coup} now becomes a Z_{coup}. For 2D wire, since $Z_{\text{2D,coup}}$ is simply coupling capacitance between two wires isolated through a dielectric, this can be expressed as

$$Z_{\text{2D,coup}} = Z_{\text{C}_{\text{wire-to-wire}}} \tag{32}$$

However, for 3D TSV, $Z_{\text{3D,coup}}$ becomes a complicated network considering the liner capacitance, substrate capacitance, and substrate resistance as in Figure 71 (b).

$$Z_{\text{3D,coup}} = Z_{\text{C}_{\text{ox}}} + (Z_{\text{R}_{\text{si}}} // Z_{\text{C}_{\text{si}}}) + Z_{\text{C}_{\text{ox}}} \tag{33}$$

$$Z_{\text{3D,coup}} = \frac{2Z_{\text{C}_{\text{ox}}}(Z_{\text{R}_{\text{si}}} + Z_{\text{C}_{\text{si}}}) + Z_{\text{R}_{\text{si}}} Z_{\text{C}_{\text{si}}}}{Z_{\text{R}_{\text{si}}} + Z_{\text{C}_{\text{si}}}} \tag{34}$$

Using Keysight ADS, $Z_{\text{3D,coup}}$ and $Z_{\text{2D,coup}}$ are compared when the pitch is 10μm (when 2D wire and 3D TSV have same dimensions) in Figure 73. Red line denotes Z_{2D}, and Blue line denotes Z_{3D}. It is shown that Z_{2D} is a linear curve since it only sees the capacitive load. However, Z_{3D} shows a non-linear curve in the GHz region because the conductive silicon substrate (R_{si}). R_{si} combined with C_{si} and C_{ox} forms a coupling network impacting in the GHz region. From Figure 73, this study reports that (1) $Z_{\text{3D,coup}}$ is always lower than $Z_{\text{2D,coup}}$ in all frequency regions. This means that the 3D timing path will suffer from more delay than in the 2D path. (2) The impedance ratio between $Z_{\text{3D,coup}}$ and $Z_{\text{2D,coup}}$ roughly leads to delay ratio between 2D and 3D. Note that the ratio between Z_{2D} and Z_{3D} is almost the same in a broad range of frequency. At 1GHz, $Z_{\text{3D,coup}} = 5.02K\Omega$ and $Z_{\text{2D,coup}} = 12.5K\Omega$, and $\text{Delay}_{\text{3D}} = 242.0ps$ and $\text{Delay}_{\text{2D}} = 110.5ps$. Within this broad spectrum, $Z_{\text{3D,coup}}$ is 2.5x more than $Z_{\text{2D,coup}}$, and this impedance difference roughly translates into the delay ratio.

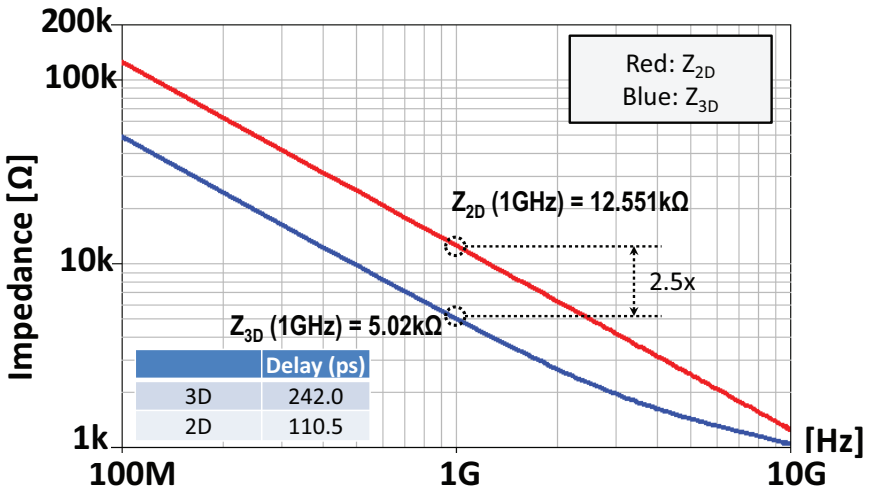


Figure 73: Coupling load impedance Z_{2D} and Z_{3D} when TSV pitch is $10\mu m$

3.7.4.3 Why 3D Delay is Not Sensitive to Neighbor Distance

This section explains why 3D delay is not sensitive to neighbor TSV distance by analyzing the impedance curve. Figure 74 (a) shows how $Z_{\mathrm{2D,coup}}$ changes when the pitch of a 2D wire changes. Note that the impedance curve of the 2D increases monotonically as the pitch increases. Thus, as the pitch increases, higher Z_{2D} leads to less timing delay. As TSV pitch increases, the coupling capacitance reduces. Lower coupling capacitance leads to less delay, which has the same meaning as the impedance analysis.

However, $Z_{\mathrm{3D,coup}}$ due to pitch increase (see Figure 74 (b)) is not as significant as in 2D. $Z_{\mathrm{3D,coup}}$ curve only changes on the high frequency region and remains almost the same below 1GHz. This is why 3D delay is not highly impacted by the increase in TSV pitch. In the equivalent model of TSVs in Figure 71 (b), the TSV pitch increase only changes the values of $Z_{\mathrm{R_{si}}}$ and $Z_{\mathrm{C_{si}}}$. Since a high-capacitance C_{ox} exists in the impedance path of $Z_{\mathrm{3D,coup}}$ as in Equation 34, the actual factor that increases $Z_{\mathrm{3D,coup}}$ in the broad frequency range is $Z_{\mathrm{C_{ox}}}$ and not $Z_{\mathrm{C_{si}}}$. $Z_{\mathrm{C_{si}}}$ and $Z_{\mathrm{R_{si}}}$ change impacts on the high frequency Z_{3D}, but this impact is not significant on reducing much delay.

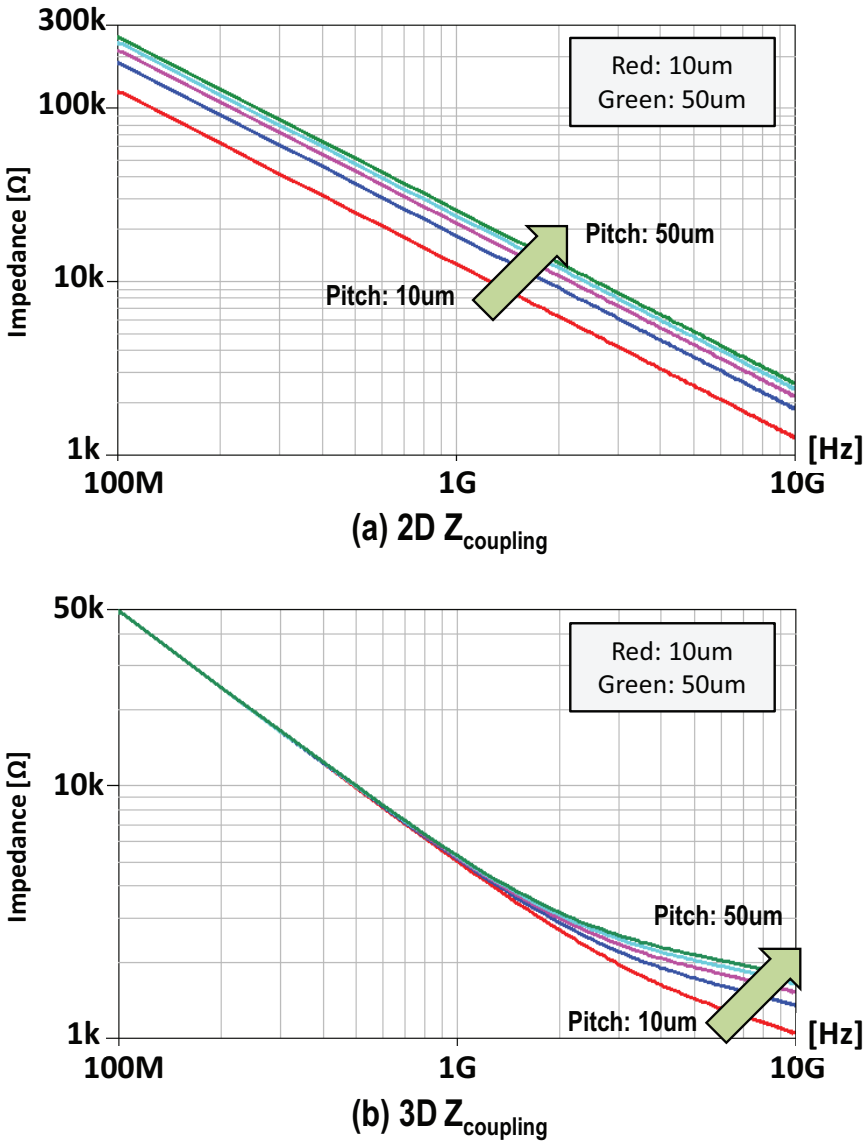


Figure 74: Z_{coup} change when TSV pitch changes from $10\mu m$ to $50\mu m$. (a): 2D, and (b): 3D.

3.7.5 Technology Impact on 3D Delay

This section studys how advanced technology nodes impact the delay on 3D TSV nets. Predictive Technology Model (PTM) 20nm, 16nm, and 10nm fin-FET transistors [63] are used and minimum sized drivers are built. Based on each technology node, VDD was scaled accordingly. TSV height varies from 20μm to 100μm to see how the 3D net delay changes. Figure 75 shows the analysis results, and through this, two important findings

are reported: (1) As technology scales, newer technology will see less delay due to TSVs. On each TSV height, delay of a 10nm 1x driver is almost 50% of 20nm 1x driver. Driver strength improves as technology scales, and this will lead to less delay in 3D TSV nets. (2) TSV is still a significant load in advanced technology nodes. Unless TSV technology (such as height, radius) scales as transistor scales, 3D net will still be a huge load to the drivers. Even when TSV height is 20um, the delay occured by TSV is more than 50ps in each 10nm, 16nm, and 20nm node with 1x driver.

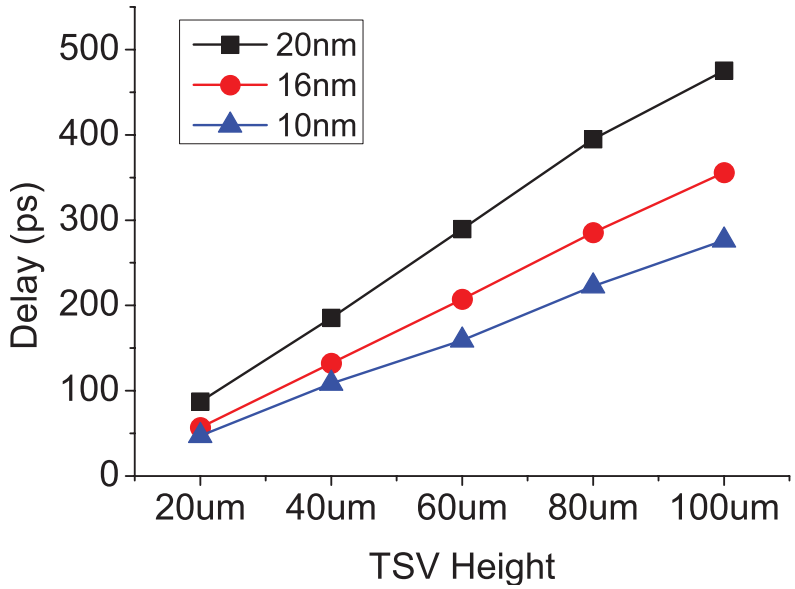


Figure 75: Delay impact when technology scales from 20nm to 10nm (driver size: 1x). TSV height scales from 20μm to 100μm.

3.7.6 Full-chip Impact on Timing and Power

Like in Sec. 3.6, this section varies the TSV parameters on the design performed on Sec. 3.5 to gain understanding of how TSV impacts the full-chip delay and power when using accurate TSV model. Table 8 shows the full-chip timing/power analysis results. From the full-chip results, the following important points are emphasized: (1) As TSV height increase, SiO_2 liner thickness decrease, and TSV radius increase, more longest path delay

(LPD) and total negative slack (TNS) are seen. This is because as these parameters increase/decrease, the total capacitance increases and leads to more timing delay. (2) Within the range of parameter change, TSV radius increase leads to the worst results in increasing the LPD and TNS. This is a similar trend to what was observed in Sec. 3.6. (3) The power increase of 3D nets due to TSV parameter change shows a similar trend to the LPD and TNS increase trend. Since the power consumption of each net is directly proportional to the capacitance increase, this is reasonable. Note that the total 3D net power increases 74% when TSV radius changes from 2μm to 10μm. Due to the TSV radius increase, not only the TSV capacitance increases, but also the number of effective aggressors to a victim increases as well.

Table 8: Full-chip timing report: Impact of TSV parameters

TSV height	20μm	40μm	60μm	80μm	100μm
LPD (ns)	2.761	2.788	2.816	2.844	2.871
TNS (ns)	-64.44	-67.32	-70.23	-73.29	-76.41
3D net power (mW)	1.72	1.75	1.78	1.81	1.83
Power increase (%)	-	1.7	3.4	5.2	6.3
Liner thickness	0.1μm	0.2μm	0.3μm	0.4μm	0.5μm
LPD (ns)	2.832	2.827	2.823	2.819	2.816
TNS (ns)	-72.45	-71.73	-71.14	-70.65	-70.23
3D net power (mW)	1.80	1.79	1.79	1.78	1.78
Power increase (%)	-	-0.6	-0.6	-1.2	-1.2
TSV radius	2μm	4μm	6μm	8μm	10μm
LPD (ns)	2.816	2.868	3.4	5.88	8.36
TNS (ns)	-70.23	-76.42	-107.9	-300.1	-492.3
3D net power (mW)	1.78	1.84	2.04	2.59	3.14
Power increase (%)	-	3.3	14.6	45.5	76.4

3.8 TSV-to-TSV Coupling Reduction

Based on the findings, a TSV-to-TSV coupling reduction method in block-level and wide-I/O design is proposed.

3.8.1 TSV Path Blocking

For a layout that has an aggressor and a victim, the capacitance of the aggressor and the additional TSV both decrease when an additional TSV is included in the design (Section 3.3.2). Thus, when a space between an aggressor and a victim exists, GND TSVs are added. The proposed coupling reduction method is called "TSV Path Blocking". By adding GND TSVs, the E-field path between the aggressor and the victim is blocked, and thus reduces the coupling capacitance. Figure 76 shows how this is applied in the layout. It may be thought that adding more TSVs will increase the total capacitance significantly. However, in a layout, a TSV is surrounded by many neighbors that the total coupling capacitance will saturate in a range around 2x (when $C_{\text{one victim}-\text{one aggressor}}$ = 1x). Thus, adding GND TSVs near the neighbor does not have a big impact on increasing the total coupling capacitance (Section 3.3.1) of a victim. The benefit of the proposed method is that, first, it recycles any empty design space in the layout so that it does not require extra silicon space just for shielding. Second, neighbor TSV does not need to be in between the aggressor and the victim for coupling reduction. E.g., assume one of the aggressors is a GND neighbor TSV in Figure 55 (b). Comparing (a) and (b), notice that the capacitance between a victim and an aggressor reduces by 23.5% (0.765x capacitance each) because two neighbor TSVs share E-field around the victim. Finally, selective coupling reduction is possible. If a victim needs more coupling reduction than other, placing more neighbor TSVs nearby helps.

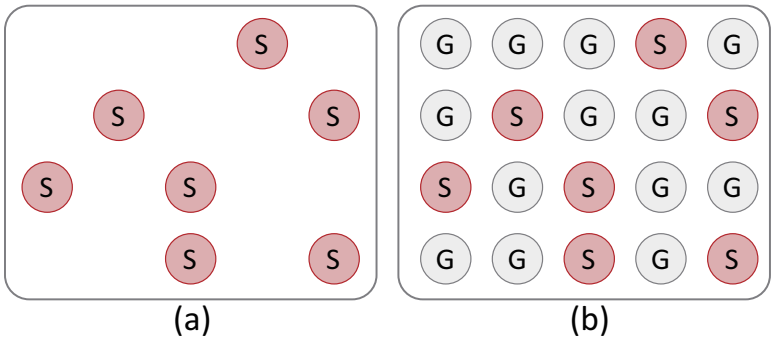


Figure 76: TSV Path Blocking in a layout: (a) Before TSV Path Blocking, (b) after TSV Path Blocking.

Table 9 shows the results. By adding TSVs inside the empty space, the total coupling noise reduces from 787V to 726V. Considering 3D noise only, the 3D coupling noise is reduced by 32% from 196V to 135V. TSV Path Blocking has a minor impact on timing. When GND TSVs are added, the total capacitance will increase slightly since more TSVs are placed near the victim. By the increased capacitive load, the total negative slack increases, but the impact is minor since the total capacitance has a maximum limit, and it is shared by the aggressor and the GND TSVs. Therefore, TSV Path Blocking is an effective way in reducing TSV-to-TSV coupling that has minor impact on timing performance.

Table 9: Impact of TSV Path Blocking - block level design

	W/O Path Blocking	W/ Path Blocking
Footprint (μm)	970 $\times$ 823	970 $\times$ 823
Total coupling noise (V)	787.42	726.04
Longest path delay (ns)	2.852	2.811
Total negative slack (ns)	-75.24	-79.62
3D coupling noise (V)	196.65	135.27

3.8.2 Optimization for Wide-I/O Design

The impact of TSV Path Blocking is shown in wide-IO design. TSV Path Blocking is an effective way to reduce coupling with the cost of increased TSV area. Three wide-I/O layouts are designed: Figure 77 (a) is the initial wide I/O design (original), (b) is the wide-I/O design with increased area (spread), and (c) is the wide-I/O design with the proposed technique applied (blocking). Figure 78 shows an actual layout applying the proposed technique, and results are shown in Table 10. For fair comparison, the placement of the modules are not changed and the area used by TSVs are only increased. If the total die size changes due to increased TSV area, the whole design will change. Thus, the die size is the same for all cases.

By the proposed technique, the TSV occupied area doubles, but the total coupling noise reduces from 824V to 742V. Considering 3D noise only, this reduces the 3D coupling noise by 45% from 193V to 105V. Note that just by spreading the wide I/O array like

Table 10: Impact of TSV Path Blocking - wide I/O design

	Original array	Spread array	W/ Path Blocking
Area by TSV (μm)	160 × 140	320 × 140	320 × 140
Total coupling noise	824.26 V	797.9 V	742.37 V
Longest path delay	2.907 ns	2.963 ns	2.925 ns
Total negative slack	-77.26 ns	-74.51 ns	-82.04 ns
3D coupling noise	193.99 V	157.41 V	105.81 V

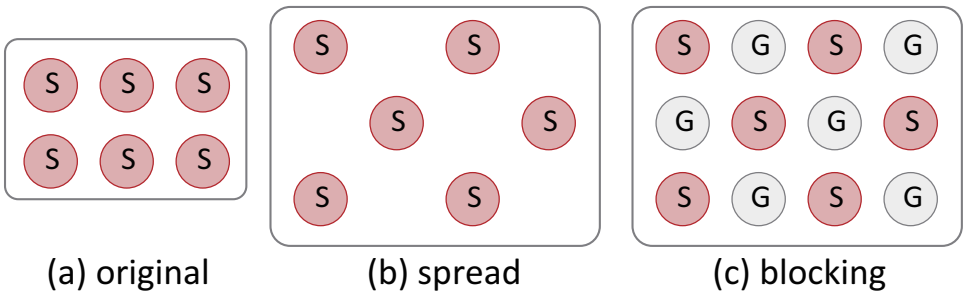


Figure 77: (a) Initial wide-I/O design (b) wide I/O design with spread TSVs (c) wide-I/O design with TSV Path Blocking

Figure 78 (d), the total coupling noise reduces too. However, if GND TSVs are included as in (b), more TSV coupling reduction will be observed. The 45% reduced 3D coupling noise would reduce the burden to the designers that requires putting significant effort to reduce 3D coupling noise using circuit techniques. E.g., wide-IO designs that consist to have complex coding scheme [39] with extra circuitry may not be needed at all due to the significant noise reduction from the proposed technique. Wide I/O with spread TSV shows less total negative slack because the capacitance that a victim sees reduces due to the increased distance. When TSV Path Blocking is applied, more coupling reduction will be observed in cost of a minor increase in total negative slack due to increased capacitance.

3.9 Summary

This chapter presented a through analysis of the TSV impact on full-chip signal integrity. First, it showed how TSV-to-TSV coupling is different in 3D ICs in comparison with package/PCB vias based on their termination conditions. Based on a realistic TSV model, this

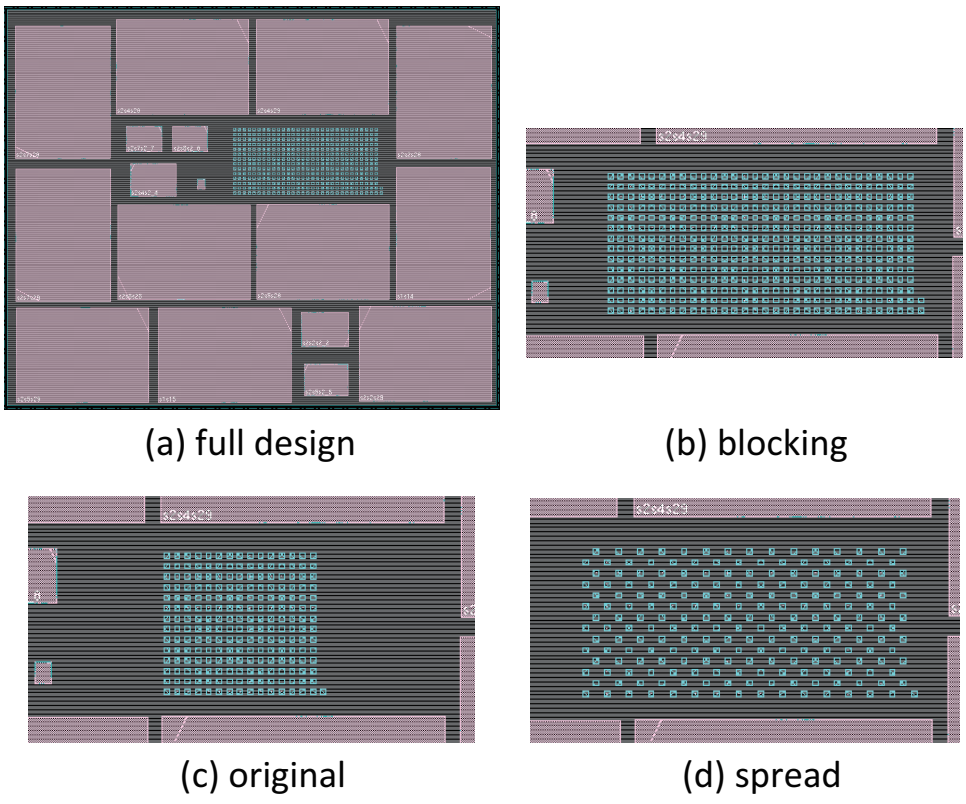


Figure 78: (a) TSV Path Blocking in Wide-I/O layout, (b) zoom-in photo of (a), (c) initial wide I/O design, (d) wide-I/O with spread TSVs

chapter analyzed the impact of port impedance on TSV-to-TSV coupling, and showed coupling is more severe in high impedance termination than in 50-Ω termination condition. Then, it was shown that TSV-to-TSV coupling has a maximum capacitance limit, and non-neighboring aggressors cause significant impact in 3D ICs, which is called the Neighbor Effect. This study developed a compact multiple TSV-to-TSV coupling model and an algorithm that accurately considers the impact of far-neighbors on full-chip 3D signal integrity analysis. Using this model, it was demonstrated that the far-neighbor aggressors have a significant impact on TSV-to-TSV coupling. Second, this study reported the TSV impact on 3D net delay. It was shown that the significant increase of the coupling coefficient did not translate to significant full-chip 3D noise increase. It was shown that 3D net delay is not highly affected by neighbor TSV distance based on the proposed "Impedance Load Analysis" method. To reduce the TSV-to-TSV coupling noise, this study proposed a technique:

TSV Path Blocking on block level and wide-I/O design. Experimental results show that by TSV path blocking, 45% 3D coupling noise reduction is achieved.

CHAPTER IV

FULL-CHIP DIE-TO-DIE PARASITIC EXTRACTION IN FACE-TO-FACE (F2F) BONDED 3D IC

Three-dimensional integrated circuits (3D ICs) have gained significant attention over the past decade as a technology that can facilitate the continuation of the advances guided by the Moores law. 3D ICs can provide significant power and performance benefits by stacking dies vertically [43]. Many studies have demonstrated the advantage of 3D ICs over conventional 2D ICs, and several companies have recently announced their plans to mass-produce commercial products based on 3D technology starting from 3D DRAMs [65]. Through-silicon vias (TSV) are one common approach for manufacturing 3D ICs. By drilling a hole inside the substrate and filling it with metal, vertical interconnections are implemented. Leveraging these vertical interconnections, shorter interconnects can be implemented, thereby leading to better performance with low power.

In addition to the TSV-based 3D IC, face-to-face (F2F) is a bonding style that also makes 3D IC possible. For 3D ICs that use TSVs, ICs are bonded by using the back side (the side where TSV is exposed) of one die and the face side (the side where top-metal is exposed) of another die. However, in F2F, the ICs are bonded by using both face sides as the bonding side using F2F bumps. Several studies indicated that F2F 3D ICs provide advantages over TSV-based 3D ICs in many applications since they do not use any silicon area [27].

Driven by the fact that scaled 3D interconnects (TSV and F2F bumps) provide denser I/Os, many studies have demonstrated how these interconnects are becoming smaller. To provide denser I/Os for F2F bonding, two technologies must scale: the F2F bump width (diameter) and distance between two dies. This is because if the distance between two

dies remain the same but the bump width scales, the bump must be manufactured to have a taller height, which would lead to reliability issues. If bump width does not scale, denser I/Os cannot be obtained. Several studies have reported bump widths in the 1μm to 5μm range [53, 50, 51]. Furthermore, F2F distances on the order of 5μm [41] and 1μm have been reported [48]. Studies have also reported direct copper-to-copper bonding that do not require any distance between dies at all [62]. Above all, these scaled F2F bonding technology proved to be reliable. Reference [40] showed that more than 3000 I/O pads were successfully bonded with these small-sized F2F bumps.

Despite the rapid scaling in F2F bonding technology, F2F bonding impact on die-to-die coupling has not been thoroughly investigated. Previous papers on F2F 3D designs extracted the parasitics of each die separately then stitched together, assuming that the impact of inter-tier coupling is not significant [27]. Therefore, this chapter first studies inter-die capacitive interactions when a 3D IC is implemented using a F2F bonding style. Using a field solver-based modeling methodology, critical aspects of capacitance in F2F bonded 3D ICs are investigated. Second, this study proposes a methodology of extracting both intra-die and inter-die parasitics in a single run on the full-chip level. Then, this study analyzes how significant the level of impact is that F2F parasitics cause. The main contributions of this work include the following:

1. Various physical and process factors are explored that affect F2F parasitics and quantify the level of error that occurs if inter-die interactions are not considered for various process and layout scenarios.

2. A holistic methodology of designing full-chip level F2F bonded 3D IC and extracting its parasitics is proposed. Using the proposed methodology, the full-chip impact of F2F parasitics is studied in various metrics.

3. It is revealed that F2F bonding causes significant inter-die capacitance and grave reduction in top-metal-to-top-metal capacitance in the same die.

4. F2F bonding causes major timing/noise error on single nets. However, the impact on the total power consumption is minor.

The results presented in this paper have important implications for both the interconnect extraction and design of F2F bonded 3D ICs with high density microbumps.

4.1 Preliminaries

4.1.1 Motivation

For dense I/Os in F2F bonded systems, smaller bump size in shorter chip-to-chip height (= H_{C2C}) is inevitable. If bump size scales but H_{C2C} does not, the aspect ratio (height) of bumps increase, causing yield problems. However, closer H_{C2C} introduces inter-die capacitance, which is significant in advanced interconnect technologies. Figure 79 illustrates the motivation of this chapter. The two boxes on the bottom (B) and the top (T) represent the top metal of the bottom tier and the top tier, respectively. All metal layers have width/spacing/thickness of 1.8/1.8/2.8μm that represents an industrial interconnect of the top metal. Synopsys Raphael is used for simulations.

In Figure 79 (a), capacitance forms only between the same tier (C_H, intra-die capacitance) because the distance between two dies is significantly large. In (b), when H_{C2C} = 10μm, inter-die capacitance C_{V1} and C_{V2} forms between tiers. Here, C_{V1} and C_{V2} are relatively small to C_H. However, in (c), when H_{C2C} becomes very close (H_{C2C} = 1μm), C_{V1} is larger than C_H (4.59fF >3.45fF), meaning that inter-tier capacitance becomes significant as H_{C2C} scales. In addition, notice that C_H reduced from 5.4fF to 3.45fF. This happens because of the E-field sharing between top and bottom-tier. When new aggressors (e.g., top-to-bottom) approach closely to the original aggressors (e.g., bottom-to-bottom) as in Figure 79 (b) and (c), E-field distributes from the original aggressors to new aggressors due to distance change. Thus, C_H reduces and C_V increases. From this, notice that (1) C_V increases as H_{C2C} scales. Especially, C_V becomes significant when H_{C2C} scales to the most advanced F2F bonding technologies (e.g., H_{C2C} = 1μm). (2) C_H reduces as H_{C2C} becomes

smaller.

Conventional (= Die-by-Die) parasitic extraction extracts the intra-die parasitics in each die then stitches them together as in Figure 80 (a) [27]. However, if Die-by-Die extraction is done in 3D designs where H_{C2C} is small, this overestimates C_H significantly. Comparing Figure 79 (a) and (c), this is 56.5%. In addition, Die-by-Die extraction cannot extract C_V that can become larger than C_H. Thus, F2F parasitics should be extracted in a holistic manner as in Figure 80 (b).

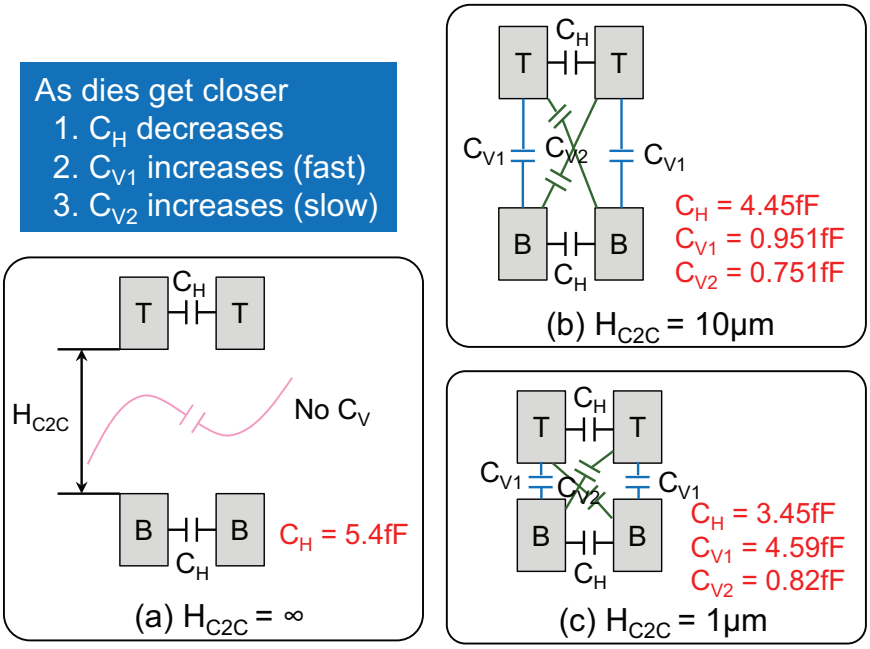


Figure 79: How capacitance changes when chip-to-chip distance changes from ∞ to 1μm. Metal dimensions: width = 1.8μm, pitch = 1.8μm, thickness: 2.8μm. C_H and C_V respectively denotes horizonal and vertical capacitances.

4.1.2 Limitations on the Top-Metal for F2F Structures

To provide meaningful results through the study, it should start with the following question: "How thick should the top metal be?" Top metals are used for various purposes such as signal routing, power delivery network design, and I/O pads for interconnection to package and PCB. When the top-metal is used for I/O pads, its thickness becomes very important. Since designed chips must go through testing, these top-metal I/O pads are the ones that

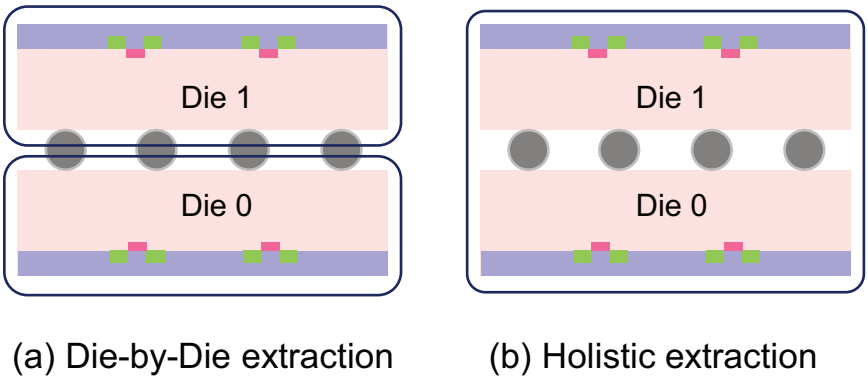


Figure 80: Two capacitance extraction methodologies: (a) Die-by-Die extraction, and (b) the proposed Holistic extraction.

are also used as probing pads during testing procedure. Note that testing probes can cause significant damage on the I/O pads. From Figure 81, it is shown that these I/O pads collapse more than 400nm after a single probe touchdown [29]. Therefore, despite the technology scaling expected on the interconnects of ICs, this chapter assumes that the top-metal will have certain limitations on the minimum thickness in order to become robust during testing. In other words, the top-metal will be assumed to be thicker than 0.6μm (400nm + margin) through out this chapter so that the top-metal do not break during testing.

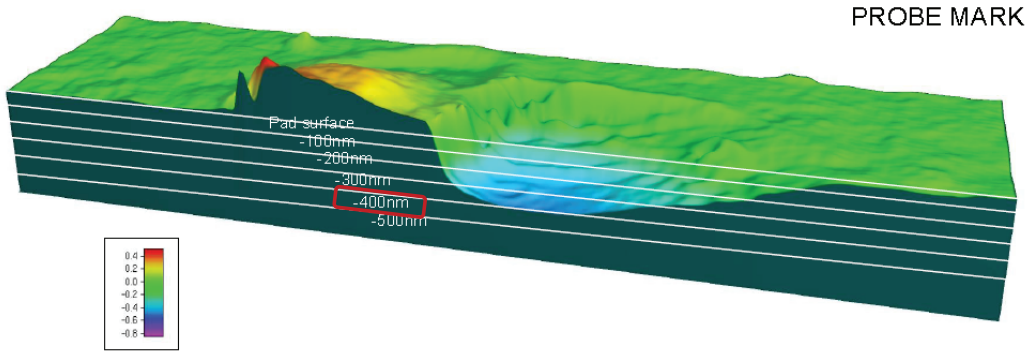


Figure 81: Damage caused to the probe pad after testing [29].

4.1.3 Top Metal Candidates

This study is performed based on a typical interconnect structure used in an industrial CMOS process technology. Table 11 shows the dimensions of the metal that is used in the

study. Figure 82 (a) shows the cross section of the top-metal interconnects in the CMOS technology, and (b) shows the cross section when two dies are stacked in F2F. Due to the damage caused to the top-metal by testing described on Section 4.1.2 (-400nm), this study is limited to see the impact of F2F coupling on the top metals that are thick enough. Thus, RDL and M9 are the top-metal candidates that are decided through out the study. From now on, RDL will be described as "thick top metal (TK)" and M9 as "thin top metal (TN)". Note that M8 is excluded as a top-metal candidate for the study because it is not thick enough. This study will refer the top-metal as "*T*" and the metal below the top metal as "*T-1*". For example, in TK case, RDL will be the top-metal (*T*), and M9 becomes the one below (*T-1*). In TN case, M9 becomes the top-metal, and M8 becomes the one below.

Table 11: Metal dimensions used in this study

	width (μm)	spacing (μm)	height (μm)
RDL (Thick top metal)	1.8	1.8	2.8
M9 (Thin top metal)	0.36	0.36	0.85
M8	0.18	0.18	0.5

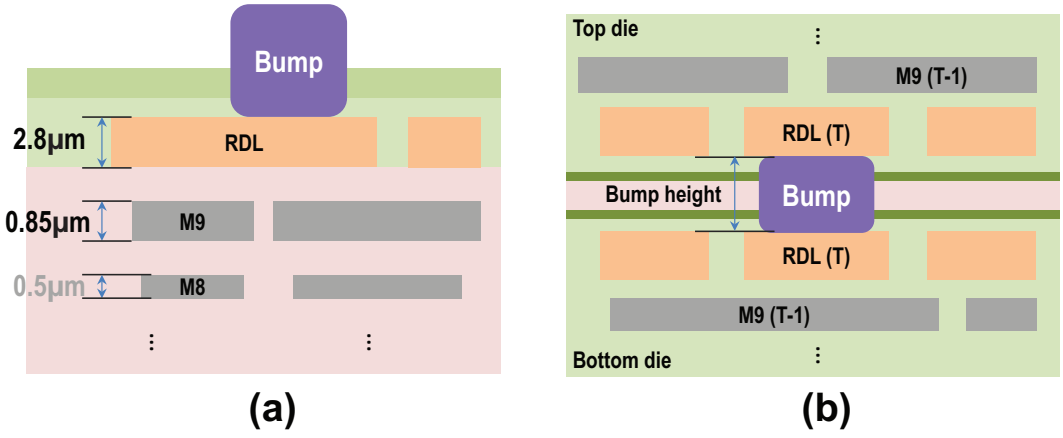


Figure 82: Interconnect structure used in this study. (a) Top metal layers in an individual die. (b) Interconnect structure when two dies are stacked in F2F 3D IC. Bump height is the distance between two dies.

4.1.4 Test Structure

When chips are stacked in F2F, the distance between metal layers is an important factor that impacts the coupling capacitance between dies. This study uses "bump height" [Figure 82 (b)] as the metric that describes the distance between metal layers. The dielectric and passivation that covers the top-metal should be open so that F2F bumps can make connection between two top-metal layers. Therefore, as these dielectrics are removed from the top-metal, the height of the bonded bumps will be the distance between top-metals in F2F stacking.

Figure 83 depicts the general 3D test structure used in the experiments in this study. Based on this test structure, it is planned to see how the coupling capacitance changes between the top metal T_0 of the top die and bottom die ($\mathrm{C_{3D}}$). To determine its significance, $\mathrm{C_{3D}}$ will be compared with the capacitance ($\mathrm{C_{2D\_1}}$ and $\mathrm{C_{2D\_2}}$) between the top metals in the same die. The length of all top metal is 40μm. *T-1* wires are placed orthogonal to T wires and are placed on its minimum pitch. *T-1* wires are long and dense enough to cover all area occupied by the top metal. By this, it is assumed that the metal layers below the top-metal are fully occupied. This models the maximum field impact from *T-1* and below so that $\mathrm{C_{3D}}$ becomes the minimum. Using this test structure, two different top-metal cases are analyzed: TK and TN.

Synopsys QuickCap NX [75] is used for the simulations. First, the model considering all details mentioned is built. Then, the capacitances are extracted from the model. Using the extracted capacitances, this study performs analysis in the following sections to examine the impact of F2F bonding.

4.2 F2F Capacitance

This section analyzes how significant F2F capacitance ($\mathrm{C_{3D}}$) is compared to the capacitance formed between metals in the same die [$\mathrm{C_{2D}} = (\mathrm{C_{2D\_1}} + \mathrm{C_{2D\_2}})/2$]. It also analyzes the

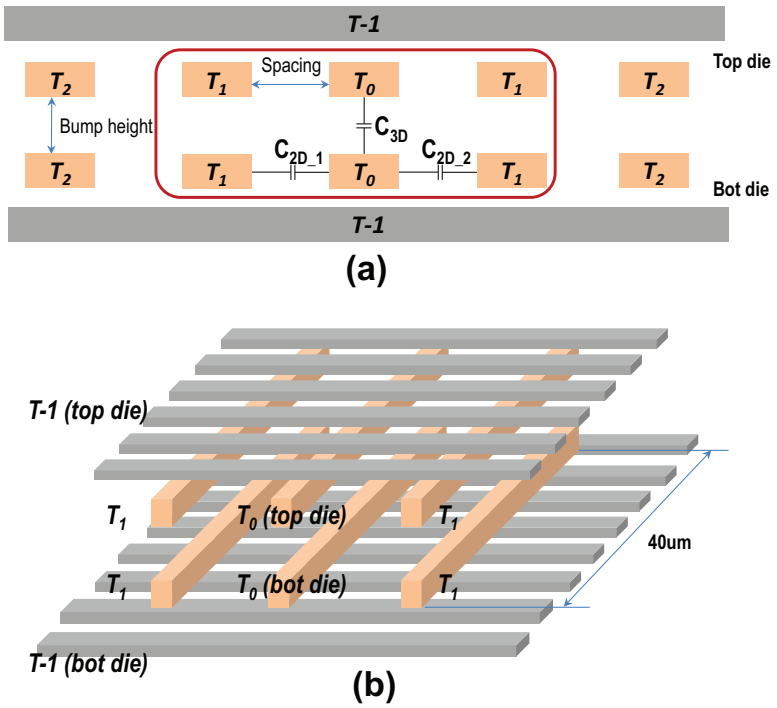


Figure 83: General test structure used in this study. (a): Cross-sectional view, (b): 3D view showing the top-metals inside the red box of (a).

factors that impact C_{3D}. Here, *3D Cap. Ratio* is defined as in Equation 35

$$3D\ Cap.\ Ratio = \frac{C_{3D}}{C_{2D}} \times 100\ [\%] \tag{35}$$

where C_{3D} and C_{2D} are the capacitances described in Figure 82 (a). "*3D Cap. Ratio* $>$ *100%*" means that the C_{3D} is bigger than C_{2D}. On the other hand, "*3D Cap. Ratio* $<$ *100%*" means that C_{2D} between wires is bigger than C_{3D} that F2F capacitance is less than C_{2D}. The following subsections first analyze the impact of F2F bonding in thick top-metal (TK) and thin top-metal (TN). Then, it analyzes other various scenarios that impact F2F capacitance in actual designs.

4.2.1 F2F Bonding Impact on Thick Top Metal (TK)

Figure 84 shows how the *3D Cap. Ratio* changes when various parameters of the top metal change: bump height, TK spacing, TK width, and TK thickness. Unless specified, the bump height and TK spacing is 5μm in all designs, and other design parameters follow

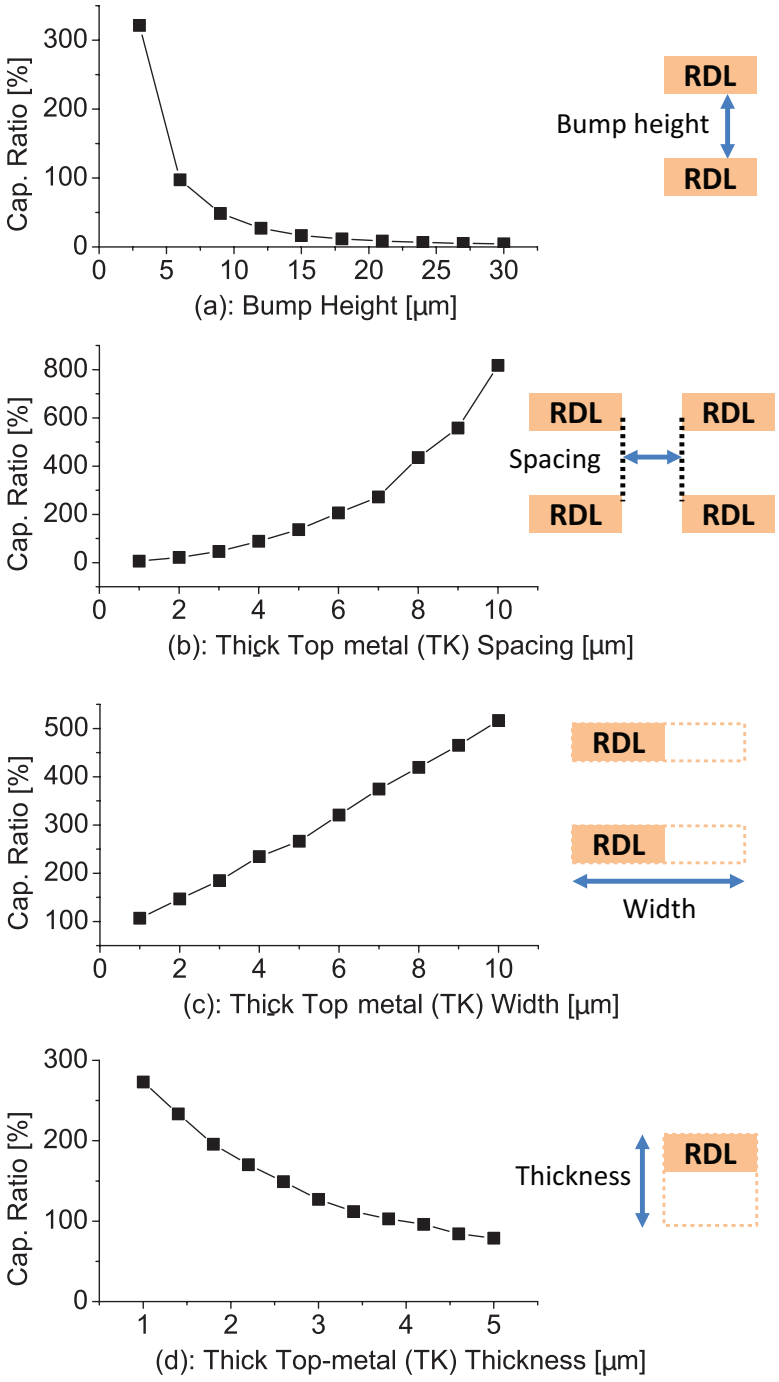


Figure 84: *3D Cap. Ratio* change due to various parameter changes in thick top-metal (TK). (a): Bump height, (b): TK spacing, (c): TK width, (d): TK thickness

the specifics of Table 11. First, (a) shows how *3D Cap. Ratio* changes when the bump height changes from 3μm to 30μm. Results show that as bump height decreases, *3D Cap. Ratio* increases significantly (321% when bump height is 3μm). This is because bump

height increase leads to both C_{2D} increase and C_{3D} reduction at the same time. In addition, when the bump height is over 10μm, *3D Cap. Ratio* becomes significantly lower than the 2D capacitance. This shows why in previous F2F bonding technologies, when the bump height was sufficiently tall (>10μm), F2F coupling was not a critical issue.

Second, Figure 84 (b) shows how *3D Cap. Ratio* changes when the spacing of TK varies from 1μm to 10μm (bump height: 5μm). Note that *3D Cap. Ratio* changes from 6.5% to more than 800% based on the top metal spacing. When spacing between top metals increase, C_{2D} reduces, but C_{3D} increases at the same time. Depending on the spacing between top metals on the same die, C_{3D} becomes significantly higher than C_{2D}.

Third, Figure 84 (c) shows how *3D Cap. Ratio* changes when the TK width changes from 1μm to 10um. As TK width increases, the *3D Cap. Ratio* increases as well. This is because when the width of the TK increases, it increases the surface capacitance between the top metals in both dies (C_{3D}). Notice that the impact of TK width on the *3D Cap. Ratio* is linear and not quadratic. C_{3D} is the only variable that changes, and TK width change has negligible impact on the change on C_{2D}. Fourth, Figure 84 (d) shows how *3D Cap. Ratio* changes when the TK thickness changes from 1μm to 5um. As TK thickness increases, a steady decrease in the *3D cap. ratio* is shown. When TK thickness increases, the capacitance between the top metal layers (C_{2D}) increase due to the increased coupling surface. However, this does not impact C_{3D} much since the coupling surface between TKs on the top and bottom die remains the same.

4.2.2 F2F Bonding Impact on Thin Top Metal (TN)

Figure 85 shows the *3D Cap. Ratio* change when various parameters of the thin top-metal (TN) changes: bump height, TN spacing, TN width, and TN thickness. Here, a more advanced bump height of 1μm is used. In addition to the bump height, the spacing between TN is fixed to be 1μm in all experiments unless specified. Figure 85 shows a similar trend as in Figure 84, but few differences occur that are unique in TN. First, Figure 85 (a) shows

that as bump height decreases, C_{3D} increases. However, notice that (1) the overall 3D capacitance ratio is smaller than in the TK case, and (2) 3D capacitance do not become bigger than 2D capacitance until the bump height is 1μm. This shows that TN will not suffer from 3D capacitance as much as TK does. Second, Figure 85 (b) shows that when the spacing in thin top-metal increases, 3D capacitance increase. However, the 3D capacitance increase ratio is more steep in TN compared to the TK case. This is because the bump height in TN is smaller than in TK. Detailed analysis regarding spacing-height relationship will be discussed in Section 4.2.3. Third, Figure 85 (c) shows how *3D Cap. Ratio* changes when the width/thickness of thin top-metal changes. Despite that exact numbers of the 3D capacitance ratio are not same as in Figure 84 (c) and (d), a similar trend is shown.

4.2.3 Spacing-Height Relationship on F2F Capacitance

Sections 4.2.1 and 4.2.2 showed a similar trend in bump height and top-metal spacing on *3D Cap. Ratio*. From this inspiration, this section studies how *3D Cap. Ratio* changes when bump height and the top-metal spacing changes at the same time. Figure 86 shows the results in thick top-metal case. It is shown from Figure 86 that *3D Cap. Ratio* is not affected by just one factor, but affected by both bump height and top-metal spacing at the same time. Note that when the bump height is the same as the metal spacing, *3D Cap. Ratio* becomes almost 1 (blue line). If the metal spacing is larger than the bump height, C_{3D} is always bigger than the C_{2D}. However, if bump height is larger than the metal spacing, C_{3D} always becomes smaller than C_{2D}. For example, when bump height is 1.4μm and TK spacing is 2.6μm, C_{3D} becomes 2.5x larger than C_{2D}. However, when bump height/TK spacing is 8.0/4.4μm, C_{3D} is only 40% of C_{2D}. Analyzing the results in Section 4.2.1, notice that *3D Cap. ratio* was almost 100% when bump height was similar to the TK spacing [see Figure 84 (b)]. Similar in Section 4.2.2, 3D cap. reaches 100% when the bump height is the same as the spacing of the thin top metal (1μm) [Figure 85 (b)].

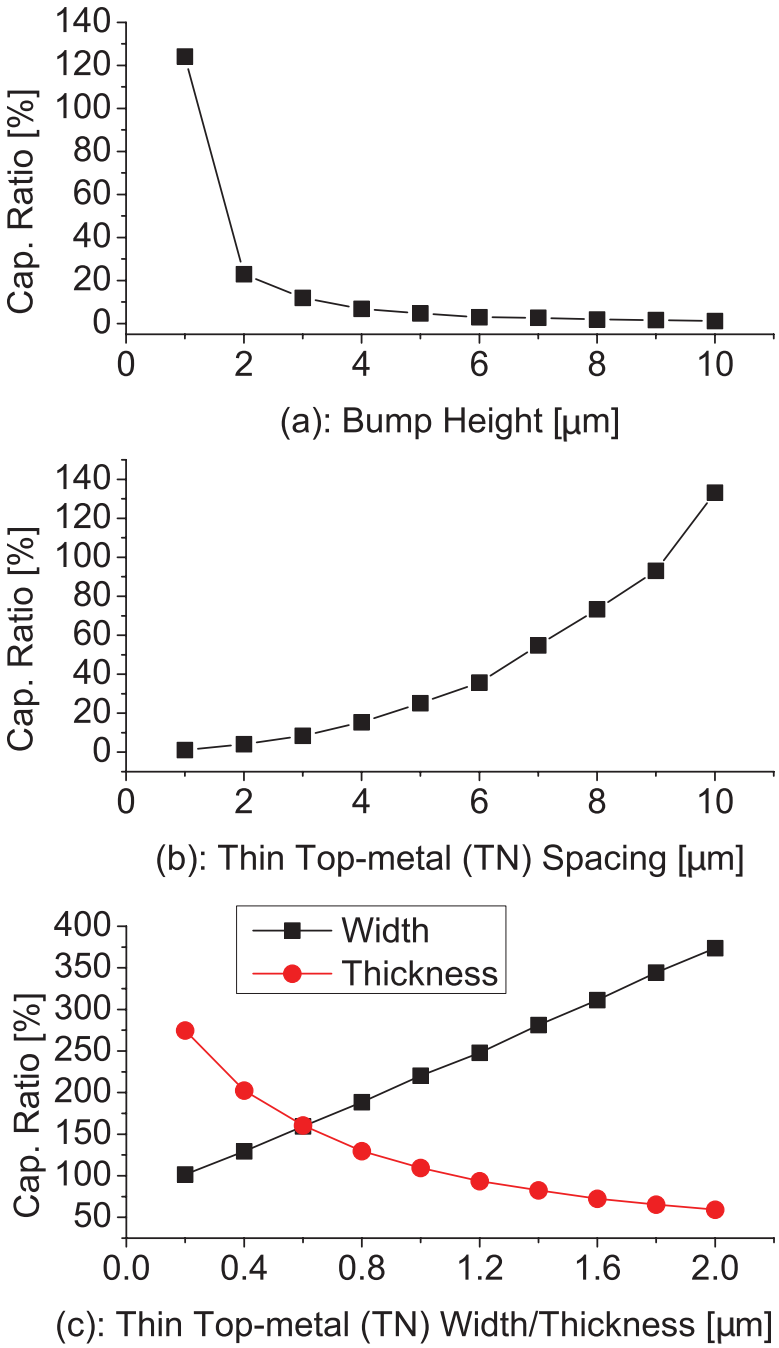


Figure 85: *3D Cap. Ratio* change due to various parameter changes in thin top-metal (TN). (a): Bump height, (b): TN spacing, (c): TN width/thickness

4.2.4 Impact of Offset Variation

Figure 87 shows how *3D Cap. Ratio* changes when the offset of the top metal changes in TK. The location of the top tier metals varies from 0 to 5um and it sees how C_{2D} and

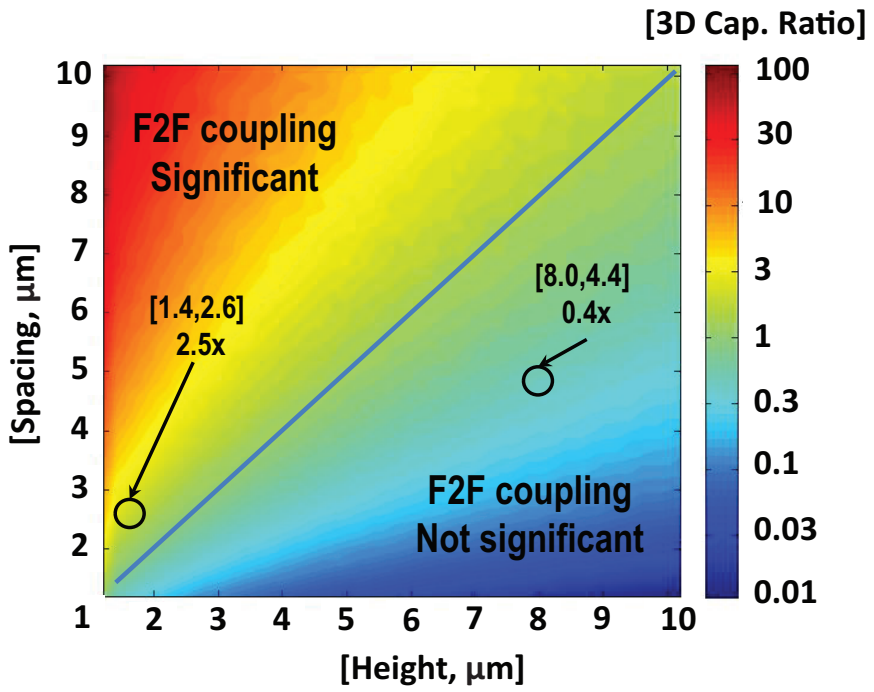


Figure 86: Impact of metal-spacing/bump-height on 3D capacitance on TK

C_{3D} changes when TK spacing/bump height are both 5μm. From the change of the offset, significant change is seen in the *3D Cap. Ratio*. Note that the change of *3D Cap. Ratio* occurs purely from the change of C_{3D} since the offset variation will not affect any change in C_{2D}. In addition, note that changing the offset of the chip will reduce C_{3D} of one top-bottom metal pair, but will increase C_{3D} formed by another top-bottom metal pair. Thus, rather than placing top and bottom tier to directly face each other, changing the offset of one tier by a few μm will reduce the 3D capacitance. However, changing the offset of a chip more one pitch will not help reducing C_{3D}. For example, if the offset is altered by exactly one pitch, the impact will be neutralized and offset changing will not do any benefit.

4.2.5 F2F Coupling in Different Top-metal Directions

The previous sections discuss the impact of coupling on F2F structures when two top metals were facing the same direction. Thus, this section examines F2F coupling when the directions of two top metals are different from one another. Figure 88 (a) shows how the test structure changes when the top-die is rotated by 90° in TK. The same dimensions are

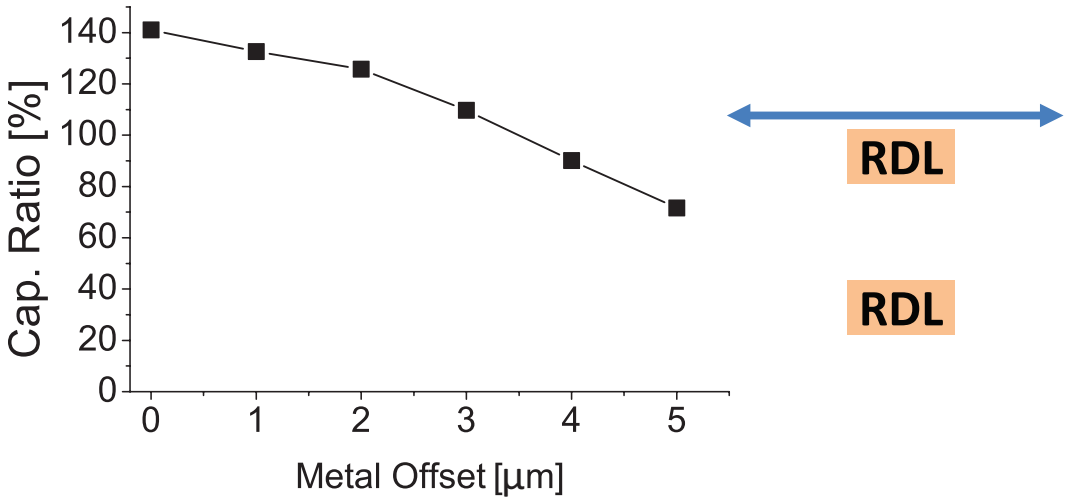


Figure 87: *3D Cap. Ratio* change when the offset of top-tier changes

used as in Section 4.2.1. Figure 88 (b) shows the extraction results in non-rotated case, and (c) shows the results in 90° rotated case. First, by rotating the top tier 90°, C_{3D} per unit metal reduces. For example, in (b), C_{3D} between the top metals is 0.899fF. However, in (c), the biggest C_{3D} between the victim and one top-tier metal is 0.259fF. Notice that C_{3D} per net reduces in 90° rotated structure. However, the total C_{3D} that a victim sees in both orientation is similar. When measuring the total C_{3D} of the bottom tier victim ["V" in Figure 88 (b) and (c)], non-rotated case gives us 1.395fF and 90° rotated case gives us 1.479fF, which the total C_{3D} is similar in both cases.

4.3 Capacitance Error Caused by F2F Bonding

Conventional parasitic extraction on F2F bonded 3D ICs normally extracts the parasitics of each die separately and stitches them together as in Figure 80 (a) [27]. This study will call this "Die-by-die Extraction". However, when the F2F bump sizes become smaller, the accuracy of the extracted capacitances in Die-by-die Extraction decreases. Therefore, extracting the F2F capacitance holistically [Figure 80 (b)] should be considered for accurate extraction. This study will call this as "Holistic Extraction". This section first reports how much error Die-by-die Extraction causes in F2F structures, and then study how the error changes due to various parameter changes.

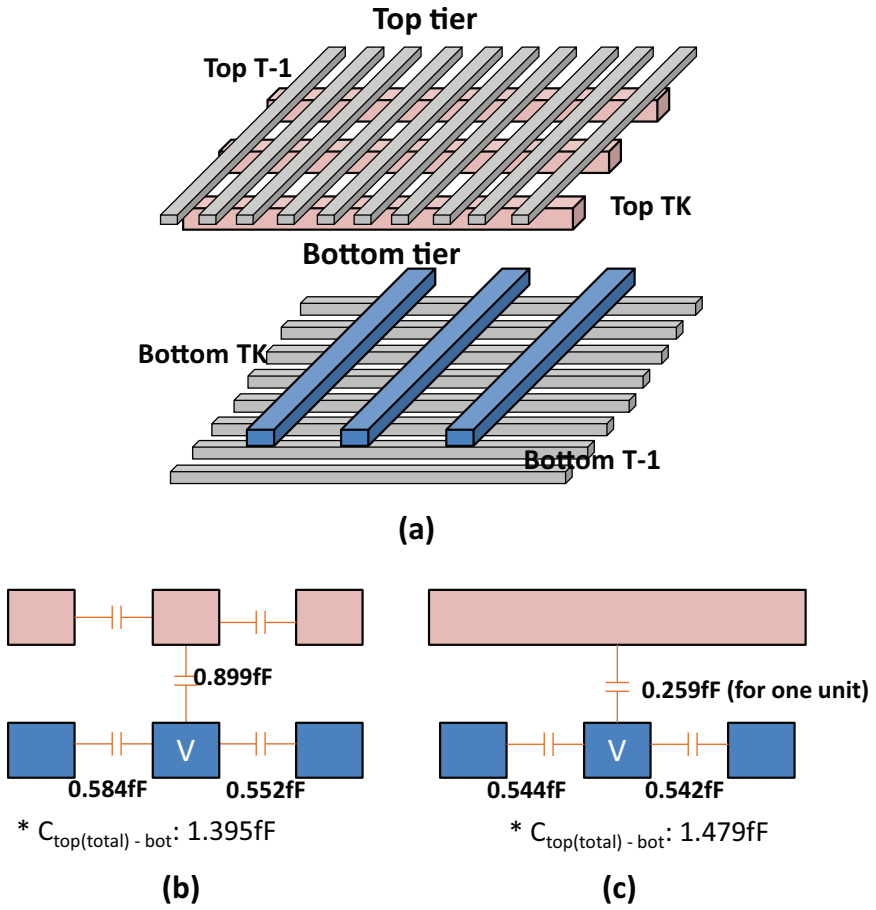


Figure 88: Top die rotated by 90°. (a) 3D view of the 90° rotated test structure. (b) Capacitance values in non-rotated structure. (c) Capacitance values in 90° rotated structure.

4.3.1 Case Studies in Different Bump Sizes

Using the same test structure as in Figure 83 (b), Table 12 shows two capacitance values in different extraction methodologies in thick top metal: (1) the total capacitance formed in the test structure, and (2) the capacitance sum of the top metal (C_{2D}). First Die-by-die Extraction is performed on the 3D structure, and the capacitance inside the whole structure is reported. Here, it obtains 10.0fF for the total capacitance, and 2.0fF for the C_{2D} formed on the top-metal layers (sum in top and bottom die). Notice that this will be the capacitance value when a 3D F2F structure is extracted in Die-by-die Extraction at any bump height. When the bump height is 5μm, however, the total capacitance is 10.2fF, and 2D top-metal capacitance is 1.3fF. This difference cannot be captured when using Die-by-die Extraction.

When the bump height is 1μm, the total capacitance becomes 15.3fF and 2D top-metal capacitance becomes 0.71fF. This means that when bump height becomes shorter, Die-by-die Extraction will cause more unwanted error. Especially, e.g., when the bump height is 1μm, the error caused will be -34.6% (Die-by-die Extraction estimates less capacitance then the correct value) in total capacitance, and 2.82x (Die-by-die Extraction estimates more capacitance then the correct value) in the top-metal capacitance. Note that as C_{3D} increases in a F2F structure, C_{2D} will see positive error since Die-by-die Extraction always overestimates, and the total capacitance will see negative error since Die-by-die Extraction always underestimates it.

Table 12: Capacitance of test structure on different bump height.

Height	Die-bydie Ext.	5μm	1μm
Total Cap. (fF)	10.0	10.2	15.3
2D Top-metal Cap. (fF)	2.0	1.3	0.71

4.3.2 F2F Bonding Impact on Capacitance Error

Figure 89 shows how the capacitance error changes due to bump height and top metal spacing on thick top metal. The baseline of this study is Holistic Extraction in the test structure, and it compares how much difference occurs in Die-by-die Extraction compared to Holistic Extraction. From Figure 89 (a), it is shown that the absolute capacitance error increases as the bump height decreases. First, significant C_{2D} error is seen when the bump height is 1μm (180.7%) and even when bump height is 10μm (22.4%). This means that Die-by-die Extraction miscalculates the capacitance between the top metals when dies become closer in F2F bonding. Second, despite the large C_{2D} error from Die-by-die Extraction, the total capacitance error is not that significant. When bump height is taller than 2μm, the total capacitance error converges to 0. This is because as C_{2D} reduces, C_{3D} increase at the same time resulting in small total capacitance difference. Therefore, the total capacitance do not change significantly. However, when the bump height becomes very small ($< 2\mu$m), C_{3D} increases faster than C_{2D} reduction. This is why the absolute total capacitance error

increases significantly in small bump heights.

Figure 89 (b) shows how the capacitance error changes when the top metal spacing change in 2μm bump height. As the top metal spacing increases, both C_{2D} and total capacitance error (absolute value) increase, because TK spacing increase in fixed bump height increases C_{3D} and decreases C_{2D}. Since the results of C_{2D} in Die-by-die Extraction assumes no obstacles over the top metal, it disregards the increase of C_{3D} due to top metal spacing. Therefore, C_{2D} error increase as the top metal spacing increase, and this also leads to the error in total capacitance [71].

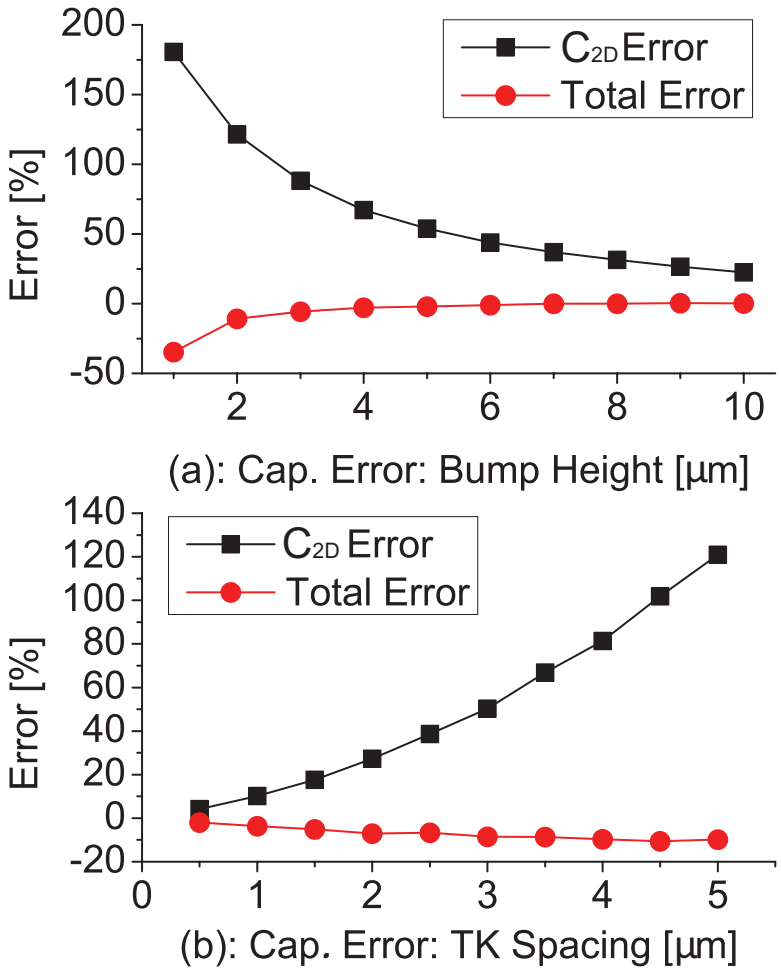


Figure 89: Capacitance error variation when using Die-by-die Extraction scheme: (a) Bump height, (b) TK spacing

4.4 Full-chip Extraction Analysis

The following sections perform full-chip level F2F study and analyze the results in two types of interconnects for an LDPC benchmark [55]. Results of other benchmarks are in Section 4.5.3. In all benchmarks, PDN [Section 4.5.2] is designed so that analysis in this study is practical.

4.4.1 Technology Setup

This study uses Synopsys 28nm as the baseline process design kit (PDK) [74]. Table 13 describes two different interconnect structures this study uses. These structures will be refered as Type 1 (Thick) and Type 2 (Thin), respectively. Both Type 1 and Type 2 consist of 6 metal layers. Type 1 uses a thick M6 width/thickness of 1.8/2.8μm and Type 2 uses M6 width/thickness of 0.36/0.85μm. For M5, each width/thickness is smaller than M6 and scaled accordingly based on the width/thickness M6 used. Note that Type 1 represents the model structure of TK, and Type 2 represents the model structure of TN in previous sections (Sec.4.2 and Sec.4.3). Note that these top-metal in both types represent the dimensions of actual industrial 28nm interconnects, and this study follows the top-metal limitation in Sec.4.1.2 so that the top metal in this study is realistic and robust during testing. For M4 to M1, this study follows the interconnects of Synopsys 28nm PDK and use the same for both in Type 1 and Type 2. For 3D stack-up, the F2F bump diameter is 1.6μm [51], and chip-to-chip distance is 1.5μm [49]. This study assumes that when a F2F design is completed in Type 1 (or Type 2), both dies will have the same Type 1 (or Type 2) interconnect structure.

4.4.2 Extraction Flow

Figure 90 proposes the extraction and analysis flow in this study. First, a 2D netlist is partitioned into two tiers and placement is done on each die. The placer in this study is based on a force-directed 3D gate-level placement engine [33], and it is modified accordingly to perform placement in the proposed F2F design flow. This gives the placement results for

Table 13: Interconnect dimensions used in this design.

	Width (μm)	Spacing (μm)	Pitch (μm)	Thickness (μm)	Dielectric (μm)
Type 1 (Thick)					
M6	1.8	1.8	3.6	2.8	-
M5	0.36	0.36	0.72	0.85	0.8
Type 2 (Regular)					
M6	0.36	0.36	0.72	0.85	-
M5	0.224	0.236	0.46	0.38	0.38
Common in Type 1 and Type 2					
M4	0.112	0.116	0.228	0.19	0.19
M3	0.056	0.056	0.152	0.095	0.09
M2	0.056	0.056	0.152	0.095	0.09
M1	0.05	0.05	0.152	0.095	0.09

the two tiers (Die0.def and Die1.def). Once the placement is done, the proposed F2F Layer Generator is used to generate a two-tier holistic F2F stack for routing and extraction. First, the proposed F2F Layer Generator assigns the standard cells on the top (Die 1) and the bottom (Die 0) of the stack by using the placement from the previous step (Die0.def and Die1.def). Second, F2F Layer Generator creates a platform that models all metal layers of both dies and the F2F interface as one holistic fashion for the interconnects. Based on the proposed platform, a holistic full-chip F2F bonded 3D design (f2f.def) can be made in Cadence SoC Encounter (A commercial P&R tool) for full-chip F2F design and impact study. Given the 3D F2F platform, Synopsys StarRC is used to extract both intra-tier and inter-tier (F2F) parasitics in just one run (.SPEF). Despite that the proposed platform is developed using 2D CAD tools, this does not harm the accuracy of the F2F extraction results because StarRC is a 3D based EM solver. As long as the correct details of the full-chip F2F design is inserted to the solver, the proposed holistic-extraction results are accurate in commercial grade. Figure 91 (a) shows an illustration of the result by the proposed F2F Layer Generator, and (b) shows a layout shot of the final result (benchmark: AES) after 3D design is completed. In detail, Figure 92 shows each metal layer, which the parasitics are extracted, in AES. Once the parasitics are extracted, timing/power library of the standard cells in each

die (Die0.lib and Die1.lib) is provided, and timing/power analysis is done using Synopsys PrimeTime.

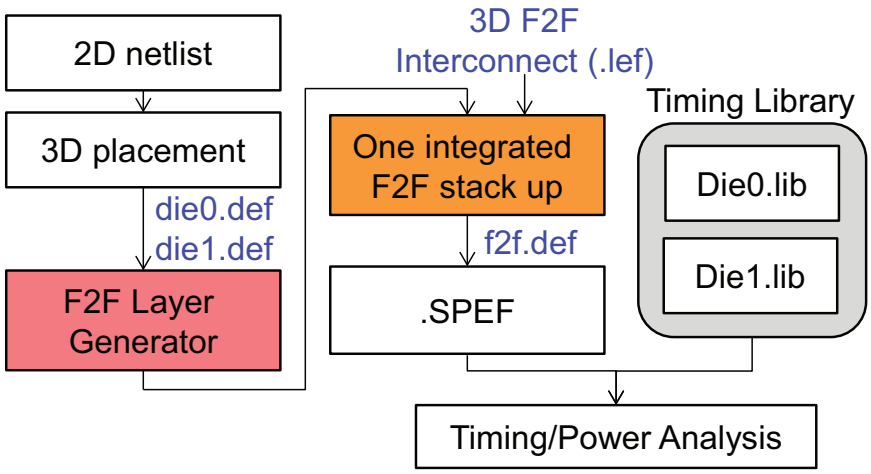


Figure 90: Proposed extraction flow using the F2F Layer Generator.

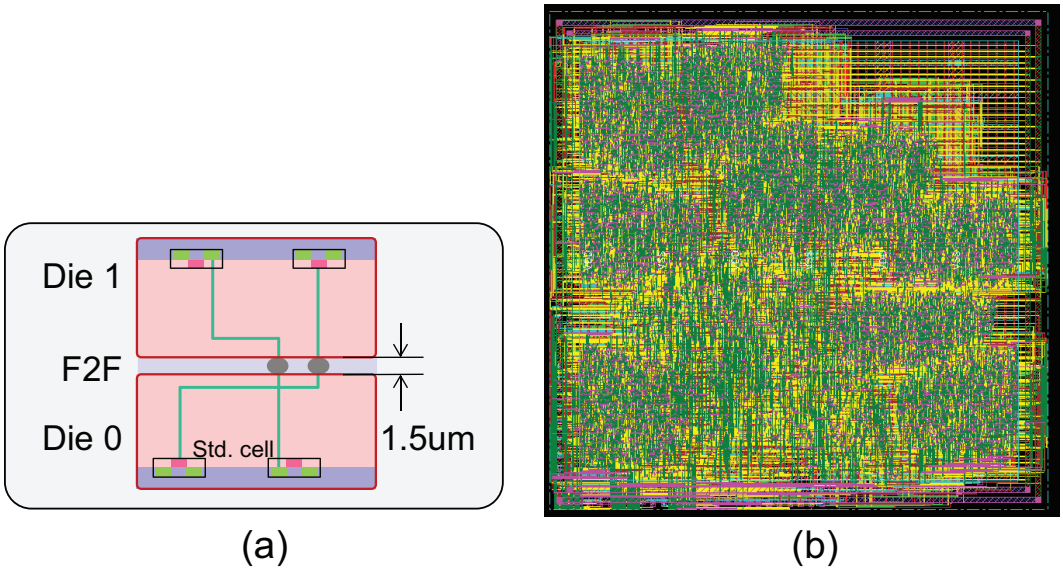


Figure 91: (a): F2F stack-up created by the F2F Layer Generator. (b): One integrated full-chip layout in Cadence Encounter with power distribution network (PDN).

4.4.3 New Capacitance in F2F Structure

This section introduces what new capacitances are formed in F2F 3D ICs. These inter-tier capacitances are defined as "F2F (3D) capacitance", and intra-tier capacitance as "2D capacitance" in this study. Figure 93 (a) shows these F2F capacitances when no bumps are

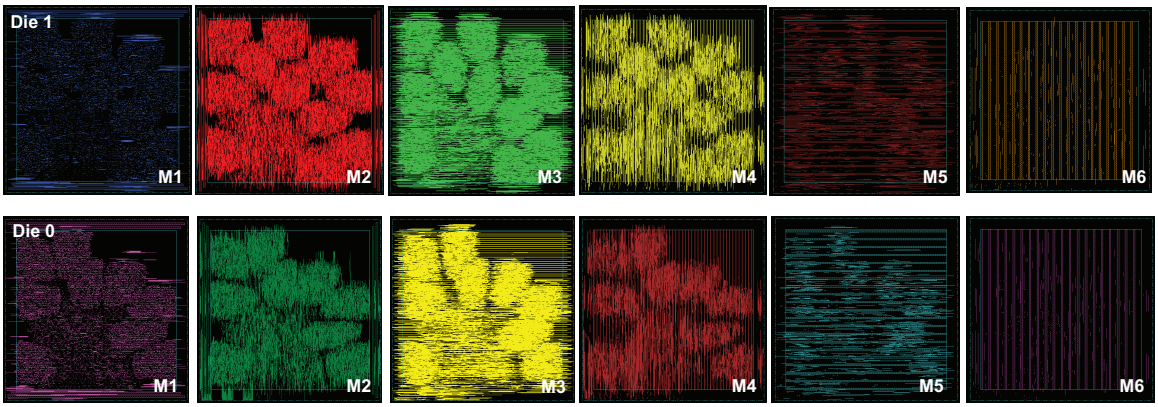


Figure 92: Individual metal layer routing in F2F implementation of AES benchmark with PDN.

between the top metals of the chip. Note that F2F capacitance are formed not only between the top metal layers (C_{F2F1}), but also between other metal layers (C_{F2F2} and C_{F2F3}). In addition, F2F capacitance not only consists of inter-metal capacitance, but also the capacitance from the bump to other structures [Bump capacitance: Figure 93 (b)]. Bump capacitance consists of two types: bump-to-bump capacitance (C_{b2b}), and metal-to-bump capacitance (C_{m2b}).

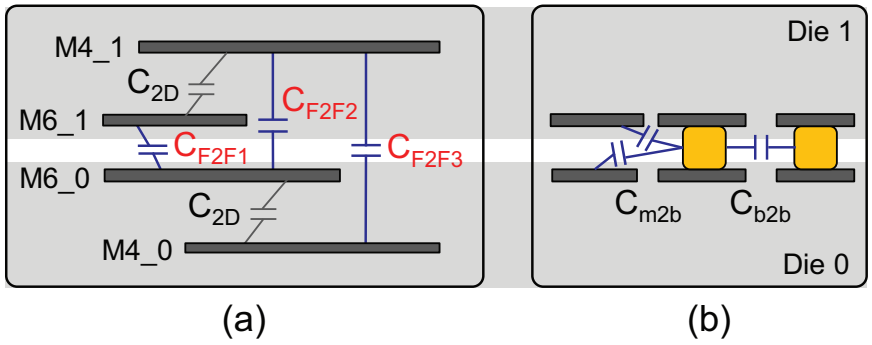


Figure 93: F2F (3D) capacitances in F2F bonding. (a): Metal-to-metal capacitance (b): Bump capacitances.

4.4.4 Comparison with Other Capacitances

This section reports how significant F2F capacitance is to other capacitances in an LDPC benchmark. To explain this, three capacitances are reported for comparison: Total coupling capacitance inside a die (= Total die cap) as in Figure 94 (a), M6-to-M6 coupling

capacitance formed inside the same die (= M6-M6 cap) as in Figure 94 (b), and total F2F capacitance formed between the two dies (F2F cap). Table 14 shows the results. The results are explained in Type 1, followed by Type 2. The total F2F capacitance is 259.17fF. Note that this is a significant value and cannot be extracted by Die-by-Die extraction.

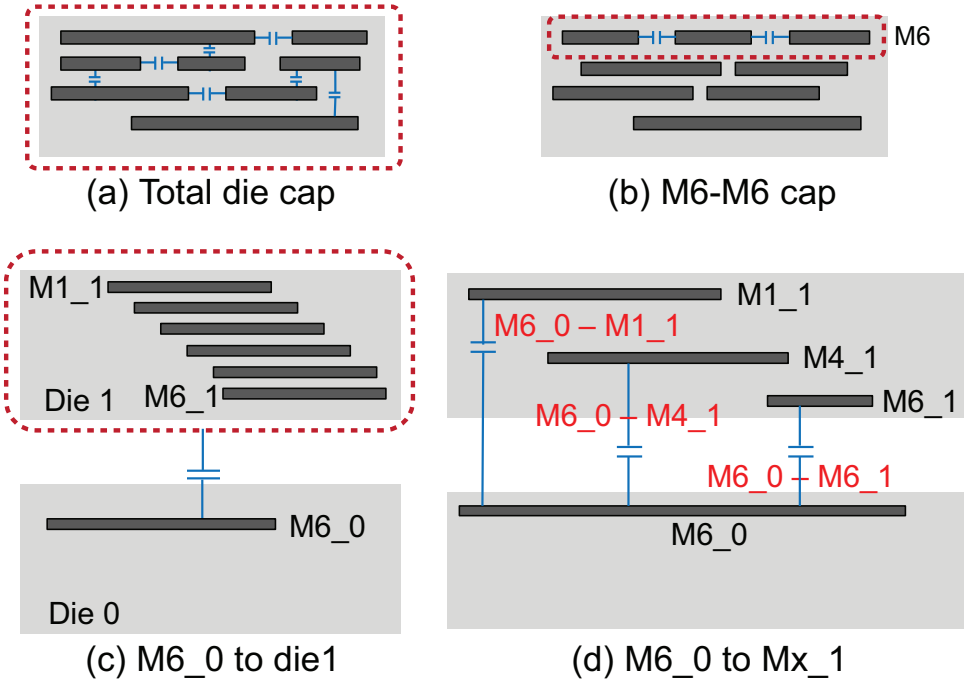


Figure 94: Parasitic capacitance definitions. (a) Total die capacitance, (b) M6-M6 capacitance, (c) M6 (Die 0) to Die 1 capacitance, (d) M6_0 (Die 0) to Mx_1 (Die 1) capacitance.

Table 14: F2F capacitance comparison to other capacitances. Total die cap and M6-M6 cap are averaged between Die 0 and Die 1. See Figure 94 (a) and (b) for definitions.

	Total die cap (fF)	M6-M6 cap (fF)	F2F cap (fF)	F2F % to M6-M6 cap	Bump cap (fF)
Type 1	38738.19	451.06	259.17	**57.5%**	116.54
Type 2	38209.93	252.11	155.49	**61.7%**	41.14

The following points are noted: (1) The total coupling capacitance formed in a die is 38738fF, and compared with this, F2F capacitance is only 0.67% of what is formed in a single die. (2) However, M6-M6 capacitance in the same die is 451.06fF. Compared with this, F2F capacitance is 57.5% of the M6-M6 capacitance. (3) Bump capacitance (116.54fF = C_{b2b} + C_{m2b}) consists of a significant portion in the F2F capacitance. A similar trend is

shown in Type 2. F2F capacitance is 0.41% of that formed in a single die, but it is 61.7% of the M6-M6 capacitance. Bump cap is also noticeable, which is 26.4% of F2F cap. The bump cap portion to the total F2F cap in Type 1 is bigger than that of Type 2. Since the metal dimensions of Type 1 is significantly larger than Type 2 (Type 1 M6 is 3.3x thicker and 5x wider than M6 in Type 2), C_{m2b} in Type 1 is bigger than Type 2. In brief, F2F capacitance contributes significantly to the total capacitance, and this impact should not be ignored.

4.4.5 F2F Capacitance Breakdown

Since the significance of F2F capacitance has been revealed, the following question remains: Between what metal layers will the most F2F capacitance be formed? To answer this question, two types of F2F capacitance breakdown is performed. First, it measures the capacitance from one metal (on Die 0) to the other die (Die 1). For example, "M6_0 – Die 1" denotes the total capacitance formed between M6 (in Die 0) and all other metal layers in die 1 [see Figure 94 (c)]. Table 15 shows that most of the F2F capacitance is formed between the top-metal (M6) to the other die in both types (98.64% in Type 1 and 97.17% in Type 2). Second, F2F capacitance is measured between each metal layers. It is shown that most of the capacitance is formed between the top metal layers of each dies (M6_0-M6_1: over 90%, see Figure 94 (d) for definitions) in both types. This makes sense because M6 is the thickest metal among all metal layers, and M6 shields the inter-tier E-field that tries to form capacitance between other metal layers. In short, most of the F2F capacitance is formed between the top metal layers in F2F configuration.

Table 15: F2F capacitance breakdown: See Figure 94 (c) for definitions.

	Type 1		Type 2	
	total cap (fF)	% to total F2F cap	total cap (fF)	% to total F2F cap
Die 0 – Die 1	259.17	-	155.49	-
M6_0 – Die 1	255.64	**98.64%**	151.09	**97.17%**
M6_0 – M6_1	252.06	**97.26%**	146.60	**94.28%**

4.4.6 Error in Die-by-Die Extraction

This section verifies the motivation from Section 4.1.1 in full-chip scale. It measures M6-M6 and M5-M5 capacitance (in the same die) and compares the two extraction methods (Die-by-Die and Holistic). Table 16 shows the results. In both Type 1 and Type 2, Die-by-Die extraction overestimates M6-M6 capacitance significantly (56.2% in Type 1 and 55.4% in Type 2) due to the inter-tier E-field sharing. Note that (1) M6 capacitance is significantly overestimated in Die-by-Die extraction when the inter-tier interaction between metals is not considered in F2F designs. In addition, when the distance between tiers becomes closer, the F2F capacitance (C_V) increases (see Figure 79) and, at the same time, the capacitance between metals in the same tier (C_H) decreases. (2) The capacitance overestimation happens significantly in M6 but not in M5. Thus, the F2F impact on M5 is almost negligible. In short, F2F bonding causes significant capacitance reduction in the top metal but almost negligible impact on the metal below.

Table 16: Capacitance overestimation in Die-by-Die extraction due to F2F cap in LDPC benchmark.

	Type 1		Type 2	
	M6-M6	M5-M5	M6-M6	M5-M5
Holistic (fF)	451.06	2890.5	252.11	1875.8
Die-by-die (fF)	702.67	2870.9	392.77	1882.2
Error (%)	56.2%	-0.7%	55.4%	0.3%

4.4.7 Impact of Chip-to-Chip Distance

Figure 95 shows how the capacitances change when the chip-to-chip distance (H_{C2C}) changes from 1μm to 10μm in LDPC benchmark both in Type 1 (a) and Type 2 (b). It also reports the change in M6-M6 capacitance in the same die. In both interconnect types, F2F capacitance converges to 0 and M6-M6 capacitance saturates to the Die-by-Die extracted value as the distance increases ($H_{C2C} = \infty$). First, in Type 1, M6-M6 capacitance reduction shows a steeper slope and starts changing more even in a far F2F distance than in Type 2. For example, when H_{C2C} = 5μm, Type 1 shows -89.1fF reduction while Type 2 shows only -12.8fF.

Comparing the two interconnect types, Type 1 M6 has wider pitch (3.6μm) than in Type 2 (0.72μm). Because of this, M6-M6 is loosely coupled to each other (than in Type 2) in terms of E-field strength. Therefore, F2F coupling starts affecting even from a far distance apart. Comparing the ratio of "F2F distance/metal pitch", Type 1 shows 1.38x (5/3.6) but Type 2 shows 6.94x. This indicates that the relative F2F distance that Type 1 sees is 5x closer than that of Type 2. This is why M6-M6 capacitance drops faster in Type 1.

Second, F2F capacitance increase in closer distance (1μm-2μm) occurs more in Type 2 (3.08x). Type 2 designs are always packed with more M6 objects than in Type 1 due to the closer metal pitch in the same area. Therefore, when chip-to-chip distance becomes closer than a certain point where its capacitance increase ratio becomes significantly high (e.g., 2μm to 1μm), Type 2 shows more F2F capacitance because it has more M6 objects than in Type 1 to form capacitance. In fact, note that when H_{C2C} = 1μm, F2F capacitance is significant in both types. This means that F2F bonded 3D ICs will suffer more from F2F capacitance in closer chip-to-chip distances.

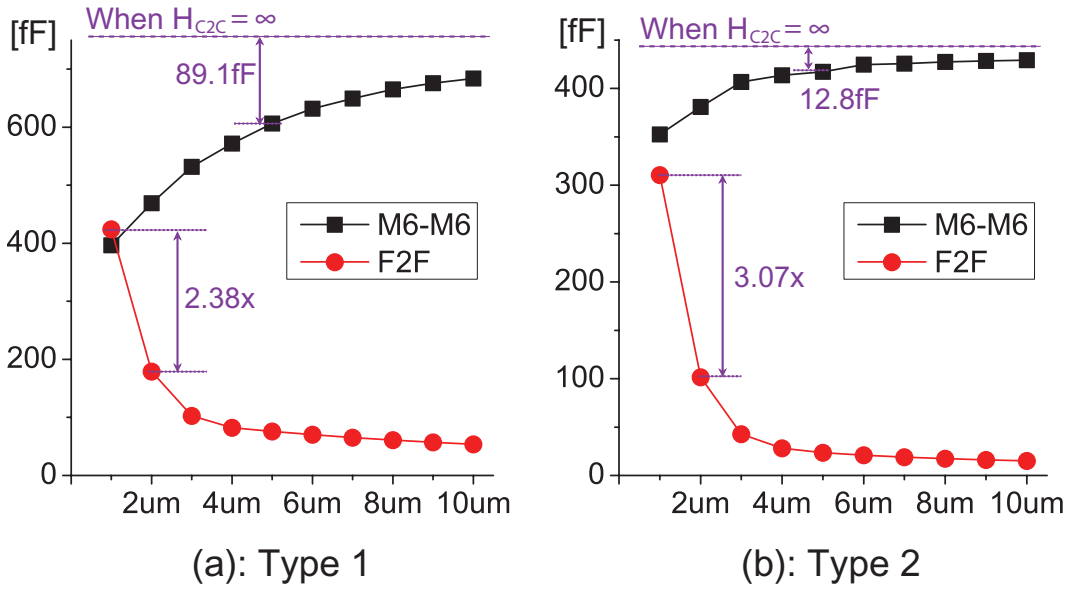


Figure 95: F2F capacitance in different chip-to-chip distance. (a) Type 1, (b) Type 2. See Table 13 for interconnect dimensions.

4.5 Full-chip Timing/Noise Impact

This section reports the impact of F2F bonding on the design such as delay, noise, and power. Compared with the Die-by-Die extraction, this section describes why Holistic extraction is necessary.

4.5.1 Holistic vs Die-by-Die Extraction

F2F bonding affect significant change in the top-metal capacitance in addition to the newly added inter-tier capacitance. Therefore, if a net uses top-metal for its routing, it will have highly inaccurate results in terms of timing and noise when parasitics are extracted by Die-by-Die extraction. Using the flow from Section 4.4.2, this study uses Synopsys PrimeTime for timing and noise analysis. It performs static timing analysis (STA) based on the clock frequency of benchmarks. Timing and noise results are analyzed in both Die-by-Die and the proposed holistic extraction, and it is compared on each net. Then, the worst case nets that show the most discrepancy in capacitance are reported.

Table 17 reports delay and noise of a net in LDPC when M6 wires are used for its routing. From this, the following important points are reported: First, in both interconnect types, Die-by-Die extraction underestimates the capacitance of a net significantly. In Type 1, F2F capacitance is underestimated by 15.67fF and this is 9.05% difference to the Holistic extraction. Because of this, Die-by-Die extraction underestimates transition time and delay by 11.02% and 18.42%, respectively. Similar in Type 2, capacitance is underestimated by 4.87fF, and, because of this, transition time and delay are both underestimated significantly. Consider a net on the critical path or a clock net uses top-metal in F2F design. These nets will see significant timing error due to underestimation in Die-by-Die extraction, which designers cannot tolerate. Second, Die-by-Die extraction leads to inaccurate noise analysis. In Type 1, the noise voltage of a net was underestimated by 50mV, and this is 83.3% noise that is missed in Die-by-Die method. In Type 2, Die-by-Die extraction does not find any effective aggressors near the victim net. However, Holistic extraction finds the inter-tier

aggressors that Die-by-Die extraction is missing and provides accurate results. In summary, Die-by-Die extracted timing/noise analysis provides highly inaccurate results due to the underestimation of inter-die (F2F) capacitance. Therefore, it is crucial to perform Holistic extraction in F2F bonded structures.

Table 17: Full-chip timing and noise analysis in LDPC benchmark.

	Holistic	Die-by-Die	Δ	Δ (%)
Type 1. Net: decoded_block_1666_				
Cap (fF)	188.902	173.232	15.67	**9.05**
Tran. time (ns)	0.150	0.141	0.014	**11.02**
Delay (ns)	0.045	0.038	0.007	**18.42**
Noise (V)	0.11	0.06	0.05	**83.33**
Type 2. Net: decoded_block_2_				
Cap (fF)	123.025	118.155	4.87	**4.12**
Tran. time (ns)	0.132	0.124	0.008	**6.45**
Delay (ns)	0.034	0.031	0.003	**9.68**
Noise (V)	0.0345	0	0.0345	NEW

The total power consumption from two different extraction methods is almost the same. For example, Type 1 LDPC consumes 49.5mW in Die-by-Die and 49.7mW in Holistic. Type 2 LDPC consumes 49.0mW in Die-by-Die and 49.1mW in Holistic. These are less than 1% difference. This is because (1) Despite the increase of F2F capacitance due to F2F bonding, the intra-die capacitance (C_H in Figure 79) also reduces at the same time. (2) In terms of the total capacitance in the full-chip, the portion that F2F capacitance contributes is very small. In addition, since M6-M6 capacitance reduces at the same time, the total capacitance difference between Die-by-Die extraction and Holistic extraction in full-chip level is almost negligible (less than 0.1% in total). Table 14 already reported that F2F capacitance to total capacitance is less than 1%. The dynamic power in digital circuits follow the equation below,

$$P_{\text{dynamic}} = CV_{DD}^2 f_{sw} \tag{36}$$

where C is the capacitance, V_{DD} is the supply voltage, and f_{sw} is the operating frequency, respectively. Since the change in total capacitance is less than 0.1% in total, which is

the only changing parameter between two extraction methods, the power difference from Die-by-Die extraction and Holistic extraction is almost negligible.

Table 18: Results for all benchmarks. M6 Δ denote the the cap overestimation caused by Die-by-Die extraction between M6-M6 in the same die as in Table 16. Cap Δ, delay Δ, and noise Δ are worst case underestimation differences caused in timing and noise analysis on a single net in Die-by-Die extraction as in Table 17.

	Area (μm x μm)	M6-M6 cap (fF)	F2F cap (fF)	F2F % to M6-M6 cap	Bump cap (fF)	M6 Δ (%)	F2F red. by PDN	Cap Δ (fF)	Delay Δ in ps	Noise Δ in mV
Type 1 interconnect (Thick)										
LDPC	700x900	451.06	259.17	57.5%	116.54	56.2%	-13.9%	15.67	7.0 (18.4%)	50.0mV (83.33%)
AES	150x150	141.67	39.18	27.7%	36.07	39.0%	-12.5%	1.91	1.6 (12.6%)	51.0mV (31.25%)
VGA	170x170	38.30	10.30	26.9%	6.82	45.8%	-19.6%	0.73	0.76 (13.2%)	32.0mV (27.27%)
JPEG	700x900	557.36	395.01	70.9%	99.35	43.6%	-19.7%	14.45	5.4 (9.2%)	137.5mV (175.0%)
M256	900x1100	353.99	371.04	104.8%	50.16	53.4%	-27.1%	10.64	13.9 (15.2%)	36.3mV (34.78%)
Avg.	-	-	-	57.56%	-	47.6%	-18.6%	-	(13.7%)	(70.33%)
Type 2 interconnect (Regular)										
LDPC	700x900	252.11	155.49	61.7%	41.14	55.4%	-50.1%	4.87	9.7 (9.6%)	34.5mV (NEW)
AES	150x150	60.81	25.91	42.6%	17.83	14.9%	-32.1%	0.51	1.1 (17.5%)	10.0mV (20.01%)
VGA	170x170	9.78	8.26	84.5%	2.52	14.3%	-42.8%	1.16	0.3 (8.8%)	10.1mV (19.99%)
JPEG	700x900	345.50	232.66	67.3%	57.05	14.8%	-52.6%	4.65	11.2 (25.4%)	40.2mV (NEW)
M256	900x1100	291.32	236.55	81.2%	24.56	26.3%	-58.3%	12.39	15.2 (5.8%)	30.0mV (28.57%)
Avg.	-	-	-	67.46%	-	25.14%	-47.1%	-	(13.4%)	(22.85%)

4.5.2 Impact of PDN on F2F Capacitance

This section proposes a F2F-aware PDN that can significantly reduce the F2F capacitance.

The PDN is formed from M6 to M3 and is designed to have a density around 20% (M6) to 10% (M3). Figure 91 (c) and (d) shows M5 and M6 PDN with signal wires. The key idea is to design a PDN on the top-metal (M6) so that it can reduce the E-field forming between the inter-tier metal layers. By having this F2F coupling aware PDN, the overall F2F capacitance reduces significantly. Table 19 reports the capacitance reduction from PDN. First, PDN reduces the total F2F capactance by -13.9% in Type 1 and -50.1% on Type 2. Type 2 interconnect demonstrates more capacitance reduction because VDD/VSS wires are placed closer to each other. Having the same PDN density among Type 1 and Type 2, more VDD/VSS wires are placed in the same unit area because M6 in Type 2 has smaller pitch and width. Note that F2F aware PDN will reduce inter-tier coupling, but it will cause power noise issues from the other die. For example, Die 0 M6 signal wires will suffer power noise from Die 0 and Die 1. This is because PDN replaces inter-tier signal-to-signal coupling capacitance into signal-to-PDN capacitance. Noise coupling between signal wires reduces by F2F aware PDN, but it causes noise coupling from the PDN of the other die. PDN also reduces inter-tier capacitance on lower metal layers (M1-M5), but note that the absolute inter-tier capacitance is already negligible even without the PDN.

Table 19: F2F capacitance reduction due to PDN.

	Type 1			Type 2		
	no PDN (fF)	PDN (fF)	Δ (%)	no PDN (fF)	PDN (fF)	Δ (%)
Tot. F2F cap	301.1	259.2	-13.9	311.4	155.5	-50.1
M6_0-M6_1	284.3	252.1	-11.3	284.1	146.6	-48.4
M6_0-M5_1	4.98	1.99	-60.0	4.04	0.70	-82.8

4.5.3 Results on Other Benchmarks

This section provides five benchmarks (including LDPC) to see the impact of F2F parasitics in various full-chip designs [55]. The biggest benchmark JPEG consists of 226K cells,

which is more than 1M transistors, and the smallest benchmark VGA consists of 5.5k cells. Benchmarks are sized optimally to perform routing without having any violations. Table 18 reports comprehensive results. Through many benchmarks, it is reported that (1) the portion of F2F capacitance to M6-M6 capacitance is significant ($>$ 67% average in Type 2), and bump cap is a big contributor to the total F2F cap. (2) Die-by-Die extraction significantly overestimates M6-M6 capacitance (M6 error, $>$ 47% average in Type 1) but not much on other layers. (3) PDN reduces F2F capacitance significantly ($>$ 47% average in Type 2). (4) Capacitance error on nets occur on full-chip designs when using Die-by-Die extraction. Due to this, the underestimated total capacitance causes significant timing (25.48%) and noise (175%) error on nets.

4.6 Summary

In this chapter, the inter-die capacitance trends were studied for various physical and process parameters when a 3D IC is implemented using a F2F bonding style. In addition, a full-chip analysis based on the proposed Holistic extraction methodology was performed on F2F-bonded 3D ICs. Based on the results, there are several general conclusions:

1. For the thick top metal layers in each die, the impact of inter-die capacitive interactions is significant when the distance between the two dies is smaller than 10 microns.

2. For the thinner metal layer below the top metal layer, the impact of inter-die capacitive interactions only becomes significant once the bump distance is smaller than 3 microns.

3. In the aforementioned process configurations, significant capacitance errors can occur when inter-die interactions are not considered in conventional parasitic extraction methods. This includes both the coupling capacitance between top metal wires in the same die (C_{2D} - overestimated due to missing inter-die shielding effects) and the coupling capacitance between top metal wires in different dies (C_{3D} - ignored).

4. Orthogonal RDL routing in facing dies can reduce inter-die coupling capacitance between individual wires. However, total capacitance and the intra-die coupling capacitance are similar to the scenario where the RDL wires are routed in parallel to the facing dies.

5. In terms of full-chip results, closer F2F distance causes significant error in M6-M6 capacitance (56.2% in LDPC) and high increase in various inter-tier capacitance that Die-by-Die extraction cannot extract (104.8% of M6-M6 in M256).

6. Among all F2F capacitances, M6_0-to-M6_1 (top metals that are facing each other) capacitance is the most significant contributor.

7. Die-by-Die extraction significantly overestimates M6-M6 capacitance (in the same die), and cannot extract accurate F2F capacitance.

8. Significant timing/noise error occurs (25.48/175%) in nets. To reduce F2F capacitance, it was found that PDN can reduce it significantly (-58.3% in M256).

These summary have important implications for both the interconnect extraction and design of F2F bonded 3D ICs with high density microbumps. Extraction tools will need to adaptively detect the distance between the two dies in a given process where inter-die capacitive interactions become significant in order to effectively balance accuracy and computational overhead. Designers and design tools may also need to consider the routing orientation of RDL layers as well as the impact of inter-die parasitics on timing, noise, and reliability in order to fully realize the potential of F2F bonded 3D ICs.

CHAPTER V

MORE POWER REDUCTION IN 3D ICS FOR MULTI-CORE PROCESSORS: THREE-TIER STRATEGIES IN CAD, DESIGN, AND BONDING SELECTION

As we reach the mobile era, power reduction is the keyword that integrated circuit (IC) industry considers as top priority. Not only for mobile devices that require long battery life and energy efficiency, but also for data centers that wish to increase their GHz/Watt performance requires to tackle this power reduction issue and have it set as their top priority goal. Power reduction directly links to packaging and cooling cost, and the power consumption of ICs has significant impact on manufacturing yield and reliability. In terms of device perspectives, the development of ultrathin body silicon-on-insulator (UTB SOI or fully-depleted SOI) and FinFET devices also correlates with this power reduction trend [4].

Due to the increasing challenges in design, power, and cost issues that industries were facing beyond 32-22nm nodes, many have started searching for alternative solutions. In this effort, three-dimensional integrated circuits (3D ICs) using through-silicon vias (TSVs) have gained a great deal of attention as a viable solution for low-power IC designs. In [7], the authors showed that -15% power reduction and +15% performance gain can be achieved by an optimized 3D floorplan in a two-tier microprocessor. In [26], authors achieved -21.2% power reduction when 3D floorplan and design techniques were applied. In [44], authors reported that -21.5% power reduction can be achieved by reducing the bus power in GPUs. In [28], authors demonstrated 50% leakage and 25% dynamic power reduction in 3D DRAM.

This chapter tries to answer the following question: ”If logic ICs are designed in many-tiers, how much more power reduction can 3D ICs achieve?” Knowing that previous 3D IC studies focused on reporting the power reduction in two-tiers [7, 26, 27, 44, 54], this chapter tries to answer the question by designing three-tier 3D ICs and studying the impact. In detail, by using an OpenSPARC T2 (a commercial multi-threaded microprocessor that has been released to public) [56] in a PDK [74] that are both available to the academic community, this study visualizes the unique design challenges and benefits of three-tier 3D ICs, which two-tier 3D ICs did not have. This study develops CAD tools for various three-tier 3D IC design styles, build GDSII-level 3D IC layouts, and perform optimization and analysis using sign-off CAD tools. The contributions of this research include the following:

1. This is the first that reported the largest power reduction that 3D ICs have. Three-tier Core results show -36% power reduction to the 2D counterpart [26] and -27.2% in full-chip, which is the biggest power reduction achieved among all other previous studies [27].

2. Three-tier 3D IC design in mixed bonding styles (e.g., face-to-face and face-to-back combined) help reduce more power. To reveal these benefits, this study develops CAD tools and implement various mixed bonding styles in three-tier.

3. Block-folding technique helps to reduce significant power in three-tier design. However, careful design management must be followed, and different bonding styles in mixed bonding impact the design quality in three-tier block-folding.

5.1 Simulation Settings

This section describes the simulation settings used in this chapter. Regarding benchmark, Section 5.1.1 describes the benchmark used in Section 5.3, 5.4, and 5.5. Benchmark used in Section 5.6 is detailed in Section 5.6.1

5.1.1 Benchmark

For the three-tier (3-tier) study, this research uses OpenSPARC T2 Core (T2 Core) [56] as the benchmark. T2 Core consists of 12 functional unit blocks including two integer execution units (EXU), a floating point and graphics unit (FGU), an instruction fetch unit (IFU), a load/store unit (LSU), and a trap logic unit (TLU). The benchmark is synthesized and designed using Synopsys 28nm PDK [74]. The PDK allows to use up to nine metal layers, and dual-V_{th} (RVT: regular V_{th} and HVT: high V_{th}) standard cells are used during the design. In addition, power distribution network (PDN) is included in the designs. A fixed PDN is placed at the initial design stage before placement and routing and is targeted to have a density of 25% (M9) to 10% (M3). Table 20 describes the details of the PDN design. This study does not place a fixed PDN for M1 and M2. This is because for M1, standard cells already contain VDD/VSS lines, and a fixed PDN for M2 acts as placement blockages.

Table 20: PDN specifications used in our 2D and 3D designs. # tracks show the maximum number of signal wires that can fit in between two adjacent P/G wires.

	Local	Intermediate	Global		
	M3	M4 - M6	M7	M8	M9
Metal width/pitch	56/152nm	112/228nm	224/456nm		
PDN density (%)	10.5	14.9	18.0	21.4	24.9
PDN width (nm)	208	340	2048		
PDN pitch (um)	1.976	2.28	11.4	9.576	8.208
# tracks	11	8	20	16	13

5.1.2 3D Bonding Technology

When stacking 3D ICs in 2-tier, two bonding styles are possible: face-to-back (F2B) and face-to-face (F2F) [see Figure 96]. In F2B bonding, TSVs are used for vertical interconnects. However, since TSVs penetrate through the silicon substrate and occupy area, using excessive TSVs lead to area overhead, which designers should avoid. On the other hand, F2F is a bonding style where it uses F2F bumps for vertical interconnects. Unlike TSVs, F2F bumps do not occupy any silicon area due to its advantageous bonding style. Table 21

summarizes the bonding-style-related settings used in this chapter. This study assumes that TSVs are much bigger than F2F bumps since manufacturing reliable sub-micron TSVs are challenging. Resistances and capacitances of the TSVs are calculated based on [31].

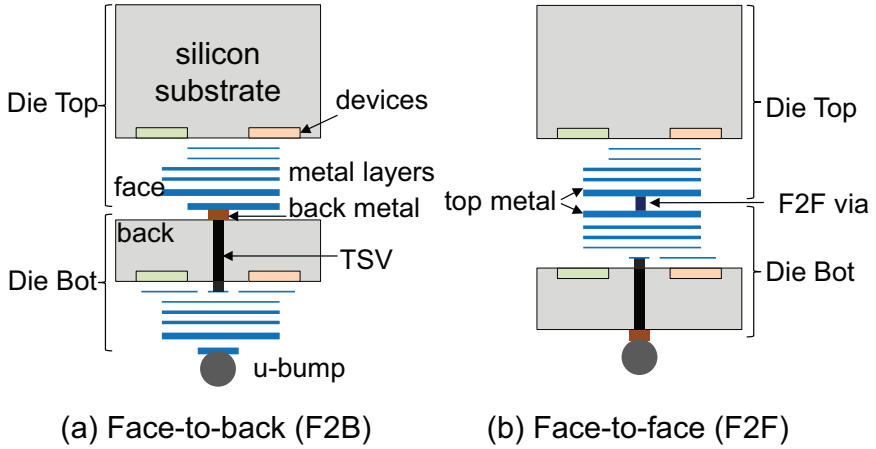


Figure 96: Basic 2-tier die bonding styles: (a) Face-to-back (F2B), and (b) Face-to-face (F2F).

Table 21: 3D interconnect settings.

	Diameter (μm)	Height (μm)	Pitch (μm)	R (Ω)	C (fF)
TSV	3	18	6	0.043	8.4
F2F bump	0.5	0.38	1	0.1	0.2

In this chapter, the impact of three different types of bonding styles in 3-tier 3D ICs are studied: face-to-back only (F2B-only), face-to-face and face-to-back combined (F2F+F2B), and back-to-back and face-to-face combined (B2B+F2F). As in Figure 97, each shows F2B-only, F2F+F2B, and B2B+F2F, respectively. In all bonding styles, Die 0 is the bottom die where it connects to the package/PCB, and Die 2 is the top die where heat sink attaches. For (a), F2B-only is a bonding style that only uses TSVs for 3D interconnects. For (b), F2F+F2B uses F2F bumps for 3D interconnects between Die 0 and Die 1, and one TSV layer (in Die 1) for Die 1 and Die 2. The advantage of F2F+F2B is that Die 0 and Die 1 suffer less from 3D interconnect penalty (smaller R and C from F2F bumps than TSVs). In addition, since F2F bumps do not occupy any silicon area and are smaller than

TSVs, more dense and optimal 3D connection can be made. For (c), B2B+F2F uses F2F bumps for Die 1 and Die 2, and two TSV layers for both Die 0 and Die 1. Since two TSV layers are used instead of one, B2B+F2F may provide less advantages than (b). However, for systems that have many external I/O connections to the package/PCB would consider B2B+F2F more beneficial than F2F+F2B. In this regard, it makes sense to use B2B+F2F bonding.

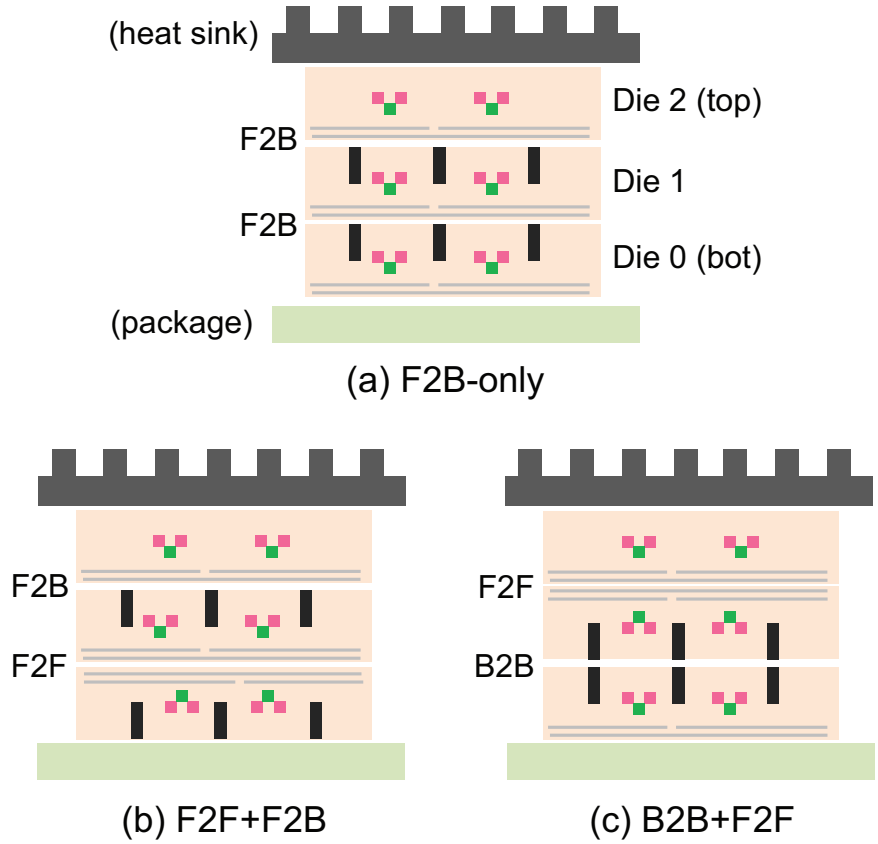


Figure 97: 3-tier die bonding styles: (a) Face-to-back only (F2B-only), (b) Face-to-face and face-to-back combined (F2F+F2B), and (c) Back-to-back and face-to-face combined (B2B+F2F).

5.2 CAD Tool for 3-Tier 3D ICs

This section first discusses existing CAD approaches for F2B and F2F 3D ICs. It also discusses why these approaches are not directly applicable to mixed bonding. Next, it describes how a 3-tier F2B+F2F mixed bonding 3D IC circuit can be constructed, and it

finally shows the modifications required to support a B2B+F2F mixed bonding 3D IC.

5.2.1 Need for New Tools

The authors of [33] have provided a framework for handling TSVs arbitrarily in a many-tier F2B-only 3D IC. However, the authors primarily compared wirelength, and when it comes to power studies, only two-tier 3D ICs have been considered in many previous papers[7, 26, 27, 44, 54].

In the placement framework proposed in [33], the gates are first partitioned into as many tiers as required. Next, TSVs are inserted into the netlist as large cells. The placement is an iterative force-directed process, with two main forces. The net force $\mathbf{F_{net}}$ tries to bring all the cells of a given net together, and the move force $\mathbf{F_{move}}$ tries to remove overlap between cells and TSVs of a given tier. The authors have also demonstrated that it is more beneficial to treat the 3D net as one subnet per tier (including the TSV), instead of as a single 3D net, as it leads to more accurate wirelength estimation. This is shown in Figure 98 (a).

When it comes to F2F integration, the placement engine remains more or less the same, with a few differences [27]. First, the F2F bumps are not inserted into the netlist, and second, the nets are not split into subnets per tier. This is because the F2F bumps are so small that they will be inserted by tricking a 2D router. Once the placement is complete, the entire 3D stack is fed into a commercial router to extract 3D via locations. However, this is limited to two tiers, with at most 7 metal layers per tier, as commercial 2D tools cannot handle more than a total of 15 metal layers.

Clearly, these approaches cannot directly be applied for a circuit with mixed bonding. TSV-based engines require TSVs to be inserted during placement, while F2F engines do not. In addition, the TSV-based engine employs net splitting, while the F2F engine does not. Finally, the F2F planner can handle at most two tiers due to commercial tool limitations. Moreover, B2B requires special handling as the TSVs in both the tiers with the B2B interface needs to be aligned. The following subsections present techniques to handle both

F2F+F2B and B2B+F2F mixed bonded 3D ICs.

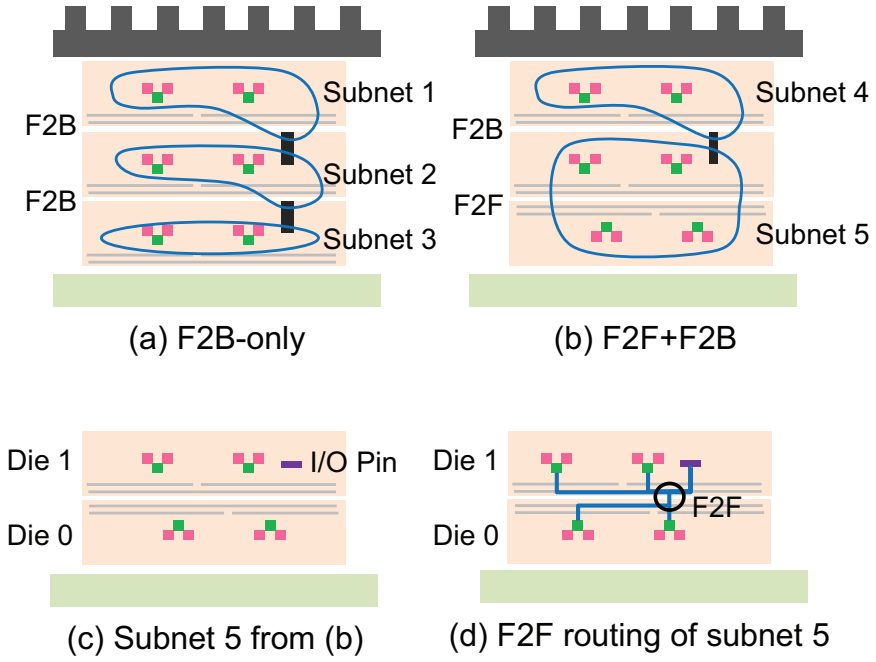


Figure 98: Net handling and routing in 3-tier mixed bonding. (a) A 6-pin net with 2 TSVs is split into one subnet per tier in F2B-only case, (b) F2F bonding does not cause net splitting, (c) Subnet 5 from (b), where the TSV is defined as an I/O pin, (d) A sample routing topology for (c).

5.2.2 CAD Tool for F2B+F2F Bonding

The modifications made to the placement engine to handle this style of mixed bonding are shown in Figure 98 (b). Two major modifications are performed. First, TSVs are inserted into the netlist only in those tiers that are F2B. Next, net splitting is performed, but do not split the nets at the F2F interface. Therefore, a 3D net spanning three tiers will have only two subnets, instead of three as in the all F2B case. Then, placement is performed to obtain the (x,y) locations of all the gates in the netlist, as well as the TSV locations for the F2B tier.

Now, F2F bumps require to be inserted using a commercial router in the F2F interface. However, as mentioned previously, commercial tools can only handle two tiers. So, the

netlist of those two tiers that are part of the F2F interface are extracted as shown in Figure 98 (c). In addition to extracting the connectivity and location of gates, additional I/O pins should be created in the same location as where the TSV would have existed. This ensures that the router will construct an accurate topology including the TSV, as shown in Figure 98 (d). Once the F2F locations are extracted, separate verilog/DEF files for each tier are created, then place, route, and optimization is performed separately.

5.2.3 CAD Tool for B2B+F2F Bonding

Handling B2B+F2F bonding is similar to the F2B+F2F mixed bonding case. Net splitting is performed at the B2B interface, and once the placement is complete, the two F2F tiers are extracted only to feed into the commercial router. The major difference is that the placer now needs to determine the location of B2B TSVs instead of a F2B TSV.

In the B2B TSV interface, both the TSVs need to be aligned. This implies that the B2B TSV can only be placed in aligned whitespace in *both* tiers of the B2B interface. First, the alignment constraint is enforced by treating the B2B TSV in both tiers as a single object with a single (x,y) location rather than two separate objects in each tier that need to be aligned. Next, the move force that removes overlap needs to consider both tiers. This is achieved by considering two move forces for this single TSV object – $\mathbf{F_{move,1}}$, and $\mathbf{F_{move,2}}$. Each force is computed separately on a per-tier basis to try and remove overlap in that tier. The aggregate move force is then the vector average of these two. Finally, once the placement is done, this B2B TSV is snapped to aligned whitespace in both tiers.

5.2.4 3-Tier 3D IC Design Flow

To design an optimized 3-tier 3D IC, First step is to synthesize the netlist with initial design constraints. Then, 3-tier floorplanning is performed using the developed mixed-bonding tools mentioned from the previous sections. Each die is designed separately based on the floorplanning results. Once the 3D CAD tools generate the TSV/F2F locations, cells and memory macros are placed using Cadence SoC Encounter. Then, the parasitics of each die

is extracted and static timing analysis is performed using Synopsys PrimeTime to obtain new timing constraints for each die. With the new timing constraints, Cadence SoC Encounter performs timing and power optimizations. Several iterations of these optimization steps (from obtaining timing constraints by Synopsys PrimeTime to design optimization in each die using Cadence SoC Encounter) are performed. By these steps, a timing-closed and power optimized design for 3-tier 3D ICs can be obtained.

5.3 Benefits of 3-Tier 3D IC

This section studies the challenges and benefits of 3-tier 3D ICs. Due to the broad scope, this section limits the study to F2B-only bonding style in block-level (non-folded) T2 Core designs.

5.3.1 New Design Challenges

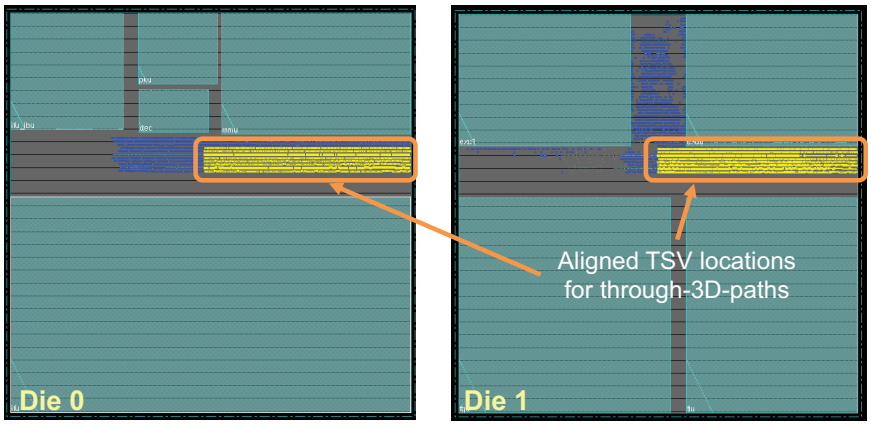
When floorplanning a 3D IC, many design constraints must be considered such as the connection between blocks and top-level pins to external connections. In addition to these constraints, area balance limits many partitioning options in a 3-tier 3D IC. For T2 Core, Table 22 shows the area ratio between the blocks inside. The two biggest modules (LSU and IFU) occupy 32.1% and 22.3% of the total T2 Core area. This means that, e.g., when a designer decides to have LSU and IFU at the same die, this die will be significantly larger than the other two since these two blocks consume more than half (54.4%) of the total area. Considering area balance, LSU should not be partitioned to be at the same die with any large blocks (such as IFU, FGU, TLU, EXU, or MMU), and the die including IFU should also be carefully be partitioned. Having this area balance issue, 3-tier partitioning becomes very challenging, and partitioning becomes even more challenging in many-tier designs.

In T2 Core, several blocks such as an LSU connect to other blocks on all three dies. If a die partition places a block (e.g., LSU) in Die 0 and the other connecting block in Die 2, Die 1 must support the paths that connect blocks in Die 0 and Die 2. These will be called as ”Through-3D-Paths.” Knowing that every block interact with other blocks in

Table 22: Area percentage of the functional unit blocks in T2 Core.

block	Area (%)	block	Area (%)
LSU	32.1	MMU	5.3
IFU	22.3	IFU_IBU	3.2
FGU	11.5	PKU	1.4
TLU	8.4	GKT	1.3
EXU0	6.3	PMU	1.3
EXU1	6.3	DEC	0.6

T2 Core, these Through-3D-Paths become as many as half of the total TSV count. Many Through-3D-Paths enter Die 1 through a TSV from Die 0 and leave Die 1 by a TSV. In this regard, Die 1 handles double the number of 3D connections than the other two tiers. Therefore, providing sufficient white space and an actual "through-path" for Through-3D-Paths is very important in 3-tier design. As in Figure 99, The white space of the top and bottom 3D connections are aligned so that these Through-3D-Paths do not need to detour. Note that the white space design in both Die 0 and Die 1 is necessary since M9 landing pads in Die 1 is on the exact location of Die 0 TSVs. If white space for Through-3D-Paths are not well designed, additional routing congestion occurs in addition to the Die 1 routing-related congestion.



(a) Die 0 (b) Die 1

Figure 99: TSV layers aligned in T2 Core to provide through path for Die 0–Die 2 connecting nets (Through-3D-Paths) in F2B-only (blue dots: regular TSVs, yellow dots: Through-3D-Path TSVs).

5.3.2 2D vs. 2-tier 3D vs. 3-tier 3D

This section now compares 2D and 3D block level T2 Core designs in TSV only bonding style. First, all designs run in a target clock period of 1.5ns (=677MHz). Note that the run speed of the designs are much slower than UltraSPARC T2, a commercial product of OpenSPARC T2, that runs at 1.4GHz [52]. This is because some custom memory blocks in T2 Core such as content-addressable memory are synthesized with cells, because a general memory compiler cannot handle these kind of memories. Unfortunately, these synthesized memories run slower than the memory macros generated by a memory compiler. Second, the baseline 2D and 2-tier 3D follow the floorplan and designs done in [26]. However, since the designs in [26] did not have PDN, PDN is included in 2D and 2-tier 3D designs and minor modifications were made to meet the timing.

Table 23 compares various metrics between 2D, 2-tier 3D, and 3-tier 3D in T2 Core designs, and Figure 106 (a) and (b) shows GDSII layouts of the 2D and 3-tier non-folded 3D design in F2B-only bonding, respectively. 2-tier 3D applies all design techniques proposed in [26]. First, by having 3-tier 3D design, the total wirelength is reduced by -36.2% and cell count by -22.8%. Compared to 2-tier 3D, this study reduces -16.6% more wirelength and -3.2% more cell count. The significant wirelength reduction comes from the smaller footprint and better top-level floorplanning.

Second, and most importantly, 3-tier 3D (non-folding) reduces the total power by -28.8%, where 2-tier 3D (block-folding) reduces -22.0% (Note that the 2-tier 3D design reduces -0.8% more power than reported in [26]). In spite of not applying block-folding in the 3-tier 3D yet, better 3-tier floorplan gives more net power reduction than in 2-tier 3D (-20.6mW more). 3-tier 3D achieves power reduction by cell count reduction, and wirelength saving. However, significant wirelength saving largely contributes to this power reduction than reduction in cell count which is not as significant (small cell and leakage power reduction). Lastly, the footprint is reduced by -67.5%. This is -14.3% more reduction than the 2-tier 3D design. In terms of silicon area, 3-tier 3D still uses -2.6% less area than

Table 23: 2D vs. 2-tier 3D vs. 3-tier 3D (non-folding, F2B-only) in T2 Core. All percentage values are with respect to 2D results.

	2D [26]	2-tier 3D [26]	3-tier 3D (non-folding)
Clock period	1.5ns	1.5ns	1.5ns
Footprint (mm^2)	**3.08**	**1.44 (-53.2%)**	**1.00 (-67.5%)**
Si. Area (mm^2)	**3.08**	**2.88 (-6.5%)**	**3.00 (-2.6%)**
Wirelength (m)	**22.4**	**18.0 (-19.6%)**	**14.3 (-36.2%)**
# Cells	523.4K	420.8K (-19.6%)	403.9K (-22.8%)
# Buffers	**221.7K**	**130.8K (-41.0%)**	**130.7K (-41.0%)**
HVT cells	370.6K	408.3K	377.4K
# TSV	-	6,562	4,118
Total power (mW)	**348.3**	**271.7 (-22.0%)**	**248.1 (-28.8%)**
Cell power (mW)	71.6	62.9 (-12.2%)	62.6 (-12.6%)
Net power (mW)	175.7	137.9 (-21.5%)	117.3 (-33.2%)
Leak. power (mW)	101.1	70.9 (-29.9%)	68.2 (-32.5%)

2D. 3-tier 3D uses more silicon area than 2-tier 3D since it requires to manage more TSVs on the top-level. However, the footprint/silicon area reduction stems from the significant wirelength and cell count reduction.

5.4 Block-Folding in 3-Tier 3D IC

This section studies how 3-tier 3D ICs reduce more power by using the "block-folding" technique. As in Section 5.3, all studies in this section are based on F2B-only bonded T2 Core.

5.4.1 3-Tier Block-Folding Challenges

Block-folding is a technique where a block inside the T2 Core is split into two (or three) tiers. Block-folding provides power reduction because it reduces the wirelength and cell count inside the blocks. In addition, it also provides better floorplanning options in the top-level. However, block-folding in 3-tier 3D must be done carefully due to its challenges. First, 3-tier folded blocks tend to have more 3D connections (use more TSVs) than in 2-tier folded blocks. Since 3-tier blocks have three partitions instead of two, this is quite obvious. In addition, area balance conflicts with minimum TSV partition. An area-balanced partition

that the designer requires will not guarantee minimum cut size for small TSV count (less area occupied by TSVs), and a minimum-TSV partition in the 3-tier will also not guarantee the desired area balance.

Second, 3-tier partition must consider external connections for TSV count management. Assume a situation where a designer should decide how to place folded IFU sub-modules when die partitioning of other blocks is done. As in Figure 100, when blue and yellow intra-IFU modules are placed on Die 2 and Die 0, the total TSV count is 48 because these two sub-modules are highly connective to each other. The TSV count doubles in this case between blue and yellow IFU modules because it must connect through Die 1. However when the blue and green IFU modules are swapped, the total TSV count reduces to 36. Though TSV count slightly increases on the inter-tier level, the total TSV count reduction becomes significant.

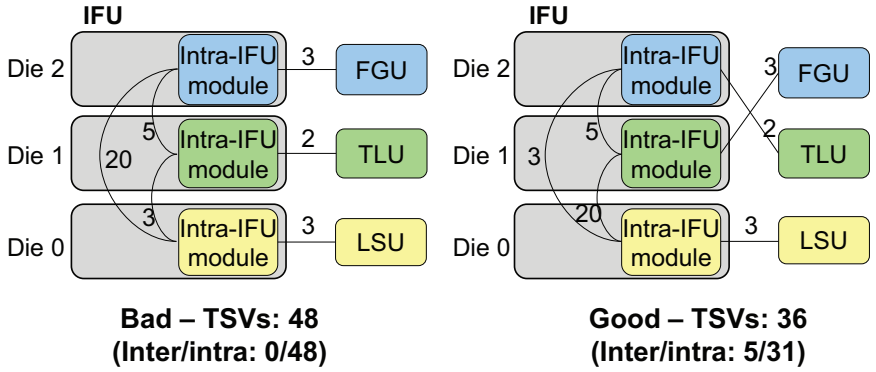


Figure 100: 3-tier IFU folding impact on intra/inter-IFU TSV count.

5.4.2 Block-Folding Strategies

5.4.2.1 Folding Blocks into 2-Tier vs 3-Tier

For 3-tier T2 Core design with block-folding, this study considers four blocks (LSU, IFU, TLU, and FGU) as candidates for folding. These modules are chosen based power consumption and average wirelength per cell so that it could give maximum power reduction. Table 24 reports cell count, wirelength, and power reduction of standalone block designs.

The partitioning and design of these blocks were done considering the top-level connections.

Table 24: Individual folded-block comparisons in F2B-only bonded T2 Core. % represents the reduction from 2D counterparts. (LSU not available for 3-tier folding)

	FGU		TLU	
	2-Tier	3-Tier	2-Tier	3-Tier
Cells	-4.6%	-6.4%	-0.3%	-0.3%
WL	-1.8%	-13.3%	-5.2%	+6.7%
TSVs	1,402	2,162	2,186	4,588
Power	**-5.7%**	**-9.1%**	**-2.9%**	**+2.1%**

	LSU		IFU	
	2-Tier	3-Tier	2-Tier	3-Tier
Cells	-4.3%	N/A	-3.8%	-5.0%
WL	-10.8%	N/A	-2.8%	-2.8%
TSVs	901	N/A	794	1,833
Power	**-7.3%**	N/A	**-1.0%**	**-1.4%**

Folding these four blocks into 2-tier gives power reduction. Each FGU, TLU, LSU, and IFU shows -5.7%, -2.9%, -7.3%, -1.0% power reduction, respectively. The power reduction stemmed from cell count and wirelength reduction. However, 3-tier folding of these blocks do not always reduce more power. 3-tier FGU showed -3.4% more power reduction than 2-tier FGU and 3-tier IFU showed only -0.4% more reduction than 2-tier IFU. Importantly, 3-tier TLU showed power *increase* (+2.1%) than the 2D TLU design. This is because TLU is highly connective between intra-TLU modules, and due to this, 3-tier TLU uses significant number of TSVs (4588 TSVs) that degrades the design quality. As shown in Figure 101, TSVs occupy a large space in 3-tier TLU in Die 0 and Die 1. For LSU, 3-tier design was not a valid option when considering TSV count, area balance, and the top-level connection.

5.4.2.2 How Many blocks Can We Fold?

It is an important decision to choose how many blocks that will be folded. Managing area balance is an important issue in non-folded designs (Section 5.3.1), and this also applies in choosing how many blocks to fold too. As in Table 22, the four candidates for folding

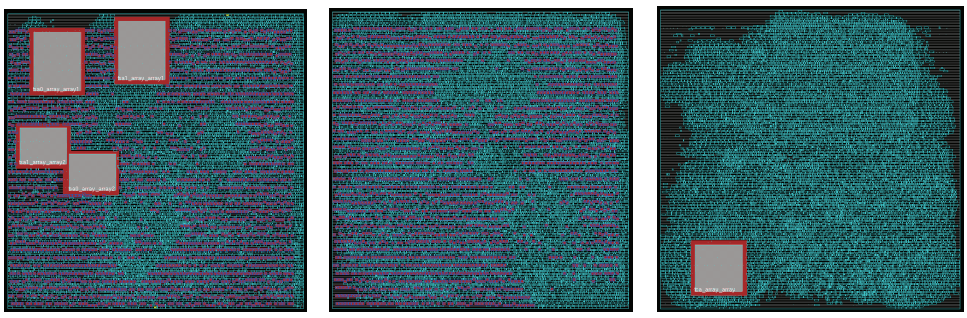


Figure 101: Many TSVs used in 3-tier TLU (in T2 Core) occupying a large area in F2B-only bonding. (purple dots: TSV)

consumes 74.3% of the total T2 Core area. However, note that in 3-tier, a folded block can be placed as Die 0-Die 1, Die 1-Die 2, or Die 0-Die 1-Die 2. No matter how it is placed, a folded block always occupy space in Die 1 [see Figure 102 (a) and (b)]. Therefore, Die 1 becomes the bottle neck when the designer needs to fold more blocks in 3-tier 3D IC. In addition, this will also conflict with floorplan options that place non-folded blocks in Die 1 because folded blocks always occupy space in Die 1.

5.4.2.3 Block-Folding For Better Floorplan

Despite that 3-tier folding for some blocks provide power reduction in the stand-alone designs, the power reduction of block-folding must be considered with top-level connectivity. Judging by the top-level connectivity and power reduction from block-folding, one good option for 3-tier T2 Core is to fold four blocks in 2-tier (IFU 2-tier). However, as Section 5.4.2.2 mentions, folding four blocks in 2-tier consumes 37.15% of the total T2 Core area. Therefore, the design footprint increases by +10%. Figure 102 (a) and (c) shows how the floorplan is done when 4 blocks are folded into 2-tiers. However, by folding IFU into 3-tiers (IFU 3-tier) [Figure 102 (b) and (c)], the die size in Die 0 and Die 1 is reduced, and the white space in Die 2 is efficiently used.

Table 25 compares the two designs (IFU 2-tier and IFU 3-tier) in various metrics. First,

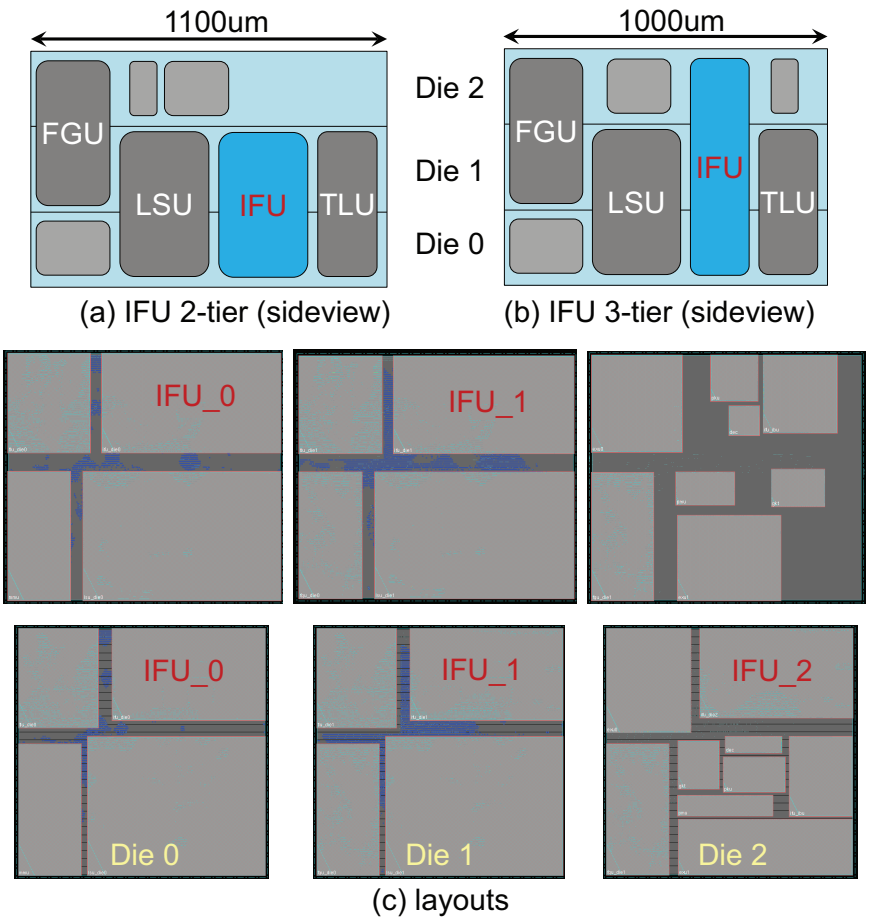


Figure 102: 2-tier vs 3-tier IFU (in T2 Core) folding impact on footprint in F2B-only bonding. (a) IFU 2-tier, (b) IFU 3-tier (footprint 10% reduced), (c) layouts.

both designs give -34% power reduction. However, IFU 3-tier shows -34.0% power reduction in addition to the 10% reduced footprint. From this it shows that block-folding can be used for better top-level floorplanning. Note that the wirelength and cell count are different in both designs. IFU 2-tier reduced more wirelength but less cell count reduction than IFU 3-tier. Different reduction ratio is shown because the commercial CAD tool optimized two different designs.

Second, comparing IFU 3-tier to non-folded 3-tier design, IFU 3-tier achieves -5.2% more power reduction than non-folded 3-tier design by careful partitioning and block-folding despite the design challenges (3-tier non-folded results in Table 23). This reveals that it is hard to predict the overall power reduction in the T2 Core just by performing

standalone design of the folded blocks. Standalone designs do not optimize the external boundaries to other blocks. Thus, top-level floorplan and block-folding benefit should both be considered for maximum power reduction. Block-folding shows better design quality in top-level too. In the top level, IFU 3-tier (block-folding) showed -32.4% cell count and -21.1% wirelength reduction compared to non-folded 3-tier. This lead to -35.8% top-level power reduction. However, note that top-level consumes less than 4% of the total T2 Core power. Therefore, the significant design quality improvement do not translate into significant power reduction. In summary, not only the standalone power reduction from the folded blocks, but also the top-level connection and area balance should be considered in 3-tier block-folding.

Table 25: IFU 2-tier vs. 3-tier in F2B-only bonded T2 Core (see Figure 102 for illustration).

	2D [26]	Block-folding (IFU 2-tier)	Block-folding IFU 3-tier
Clock period	1.5ns	1.5ns	1.5ns
Footprint (mm^2)	3.08	**1.10 (-64.2%)**	**1.00 (-67.5%)**
Si. Area (mm^2)	3.08	**3.30 (+7.1%)**	**3.00 (-2.6%)**
Wirelength (m)	22.4	12.9 (-42.4%)	13.4 (-40.2%)
# Cells	523.4K	382.1K (-27.0%)	370.9K (-29.1%)
# Buffers	221.7K	119.0K (-46.3%)	117.8K (-46.9%)
HVT cells	370.6K	358.4K	348.6K
# TSV	-	8,248	8,688
Total power (mW)	**348.3**	**230.0 (-33.9%)**	**229.7 (-34.0%)**
Cell power (mW)	71.6	57.9 (-19.1%)	54.1 (-24.4%)
Net power (mW)	175.7	103.8 (-40.9%)	107.7 (-38.7%)
Leak. power (mW)	101.1	68.2 (-32.5%)	67.9 (-32.8%)

5.5 Bonding Style Impact Study

Previous sections showed 3-tier designs in F2B-only (TSV) bonding. Thus, this section studies how various 3-tier bonding styles described in Section 5.1.2 enhance design quality and reduce power in T2 Core.

5.5.1 Bonding Impact On Floorplan

5.5.1.1 F2B-only vs. F2F+F2B Bonding

As described in Section 5.1.2, F2F bonding provides many advantages over the F2B bonding. Even in 2-tier 3D ICs, F2F reduces more power than F2B-only bonding style. Thus, it is advantageous to use F2F bonding in 3-tier designs too. However, if one layer is bonded in F2F style, the other 3D layer must be designed in F2B as bonding style. Therefore, having non-folded F2B-only T2 Core as the baseline, this study compares how the top-level design quality changes when F2F+F2B bonding is applied in 3-tier.

Figure 103 compares how the top-level design changes in Die 0 of T2 Core in F2F+F2B bonding. Note that the floorplan is exactly the same in both designs. First, F2F placement quality is much better than that of the TSV placement. Many top-level 3D connections form between Die 0 and Die 1 (2176 TSV/F2F bumps), and placing 2176 TSVs consume a large space due to the relatively large TSV size. In addition, TSV landing pads in Die 1 must not overlap with the top-metal PDN. In this regard, placing 2176 TSVs on the top-level requires more space than before. This forces the TSVs to be placed on sub-optimal locations. As in Figure 103 (a), TSVs are crowded and their locations become sub-optimal. However, since F2F bumps occupy smaller footprint than TSVs, F2F bumps can be placed on its optimal location and become less affected by the PDN. Second, because of the better F2F bump locations and small RC parasitics, top-level design quality in F2F bonding improves significantly. In Die 0, wirelength reduces by -31.9% and buffer count reduces by -39.3%. This translates to -54.5% top-level power reduction than F2B-only in Die 0.

5.5.1.2 F2F+F2B vs. B2B+F2F Bonding

For various reasons, B2B+F2F bonding can be chosen over F2B-only or F2F+F2B bonding. The difference between F2F+F2B bonding and B2B+F2F bonding lies on the second 3D interconnect layer [see Figure 97 (b) and (c)]. However, in B2B+F2F bonding style, TSVs must be placed at the same location in Die 0 and Die 1. Depending on designs, the initial

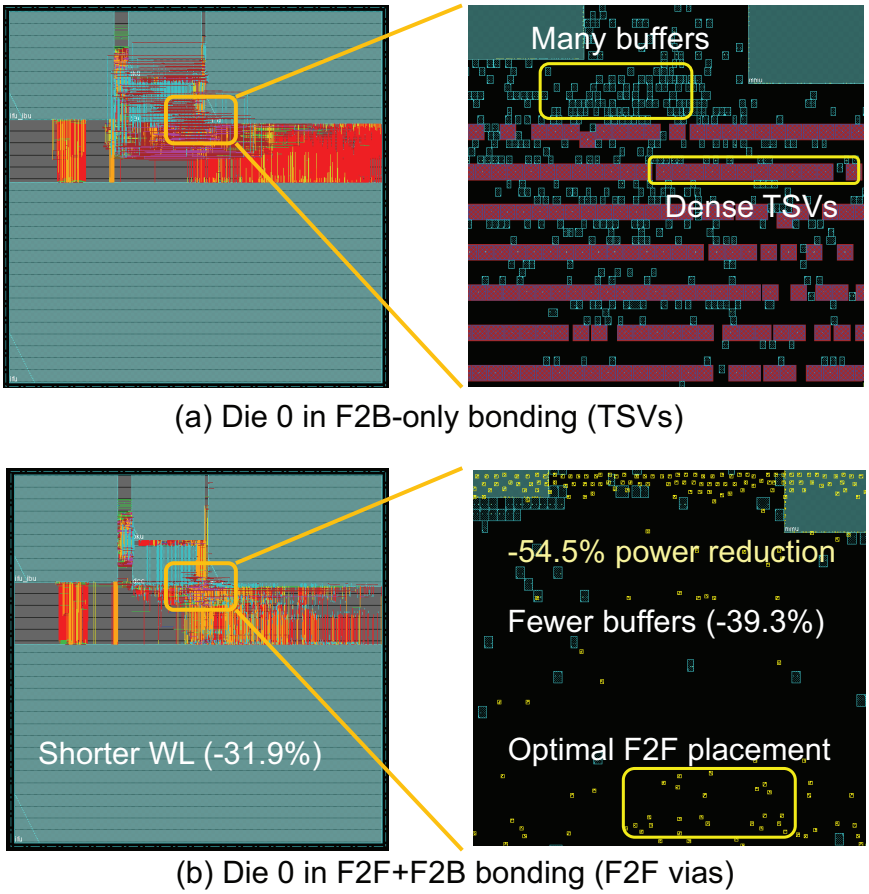


Figure 103: F2F bumps for better design in F2F+F2B bonding under the same floorplan in T2 Core: (a) F2B-only (TSVs for 3D connection), (b) F2F+F2B (F2F bumps for 3D connection).

floorplan may not align whitespace on both dies. In addition, TSV parasitics double in B2B+F2F because it uses two TSVs for 3D connection instead of one.

Figure 104 illustrates the design changes on Die 1 of T2 Core in the B2B+F2F example compared with F2F+F2B. F2F+F2B and B2B+F2F has the same floorplan, but Die 0 and Die 2 are swapped to utilize the F2F bonding for layer with more 3D connection. Figure 104 (b) shows that EXU changed its aspect ratio to provide white space for the top-level TSVs. LSU in Die 0 occupies significant area, and this forces the TSVs in Die 0 and Die 1 to be placed on the top of the layout. However, due to this, Die 1 in B2B+F2F bonding could not provide a through-3D-path because the white space between Die 0, Die 1, and Die 2 cannot be aligned. Comparing the top-level design in Die 1 (B2B+F2F vs. F2F+F2B),

the buffer count increases by +10.7% and wirelength increases by +14.3% in B2B+F2F design. In terms of the top-level power, this is +22.0% increase than the F2F+F2B in Die 1.

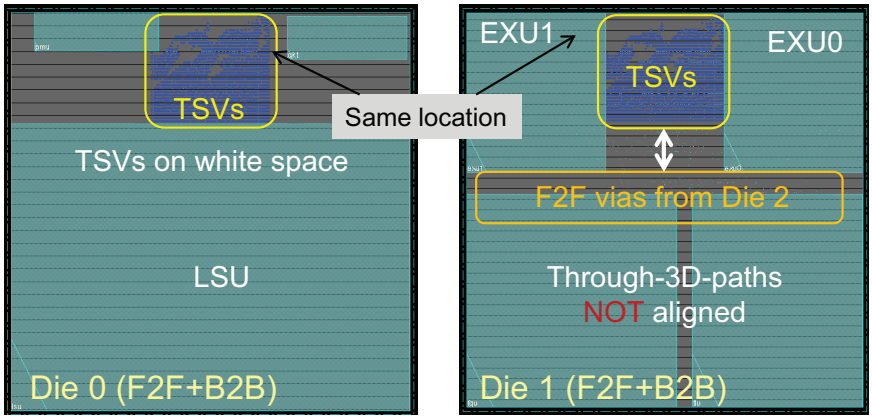


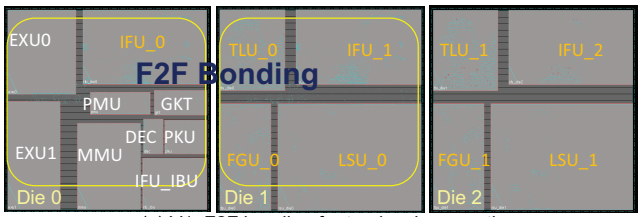
Figure 104: Through-3D-paths between Die 1 TSV and Die 2 F2F bumps not aligned in B2B+F2F bonded T2 Core because TSVs must be placed both in Die 1 and Die 2 (see Figure 99 for comparison).

5.5.2 Bonding Impact On Block-Folding

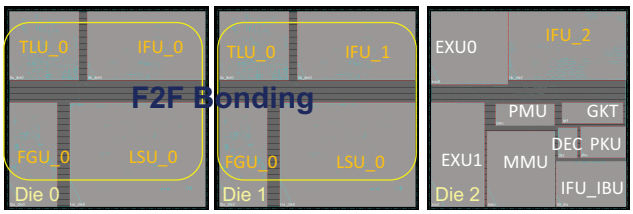
5.5.2.1 F2F+F2B Bonding on Folded Blocks

Block-folding in mixed bonding leaves the designer to choose the right 3D bonding for the right purpose. In a 2-tier design when the bonding style is decided to be F2F (or F2B), this means that both folded blocks and the top-level design utilize F2F layer. However, in 3-tier designs, designer must decide how to utilize its F2F layer since it can have only one due to the bonding technology. The more the designer chooses to use F2F layer for block-folding, the less it can be used for top-level design, and vice versa. To study which is more beneficial in T2 Core, two floorplans are studied: (1) Using F2F layer for top-level design (F2F+F2B V1), and (2) use F2F layer for block-folding (F2F+F2B V2) [see Figure 105].

The results show that F2F+F2B V1 reduces more power than F2F+F2B V2. F2F+F2B V1 showed -36.0% power reduction, but F2F+F2B V2 showed -34.7% power reduction



(a) V1: F2F bonding for top-level connection



(b) V2: F2F bonding for block-folding

Figure 105: F2F bonding choice for more power reduction in F2F+F2B bonded T2 Core. (a) F2F bonding for top-level, (b) F2F bonding for block-folding (folded blocks in orange font).

than 2D. This is explained through the following reasons: First, extra power reduction from F2F bonding in folded blocks is not significant. Block-folding based 3-tier designs must consider (1) power reduction of the block itself from block-folding, and (2) options for better connectivity in the top level. For power reduction of single blocks by block-folding in standalone designs, the total power reduction from F2F bonding is only -5.3mW. This is -1.5% of the total T2 Core power. Note that significant power reduction is not seen from folded blocks in F2F bonding. This is because 3-tier floorplanning limits many partitioning options for block-folding in F2F.

Second, top-level design quality in F2F+F2B V1 is better than F2F+F2B V2. F2F+F2B V1 and V2 uses 52% more top-level 3D connections (TSV count: 2,573) than F2B-only–block-folding design for top-level connection (TSV count: 1,693). However, since the optimal white spaces for TSV location are limited, this leads to worse TSV locations and design quality in the top level. In fact, the top-level design quality in V2 is worse than F2B-only–block-folding design. However, note that F2F+F2B V1 uses F2F layer for top-level design. Despite the increased top-level F2F bump count than F2B-only–block-folding design, F2F+F2B V1 provides better top-level design quality, and provides more

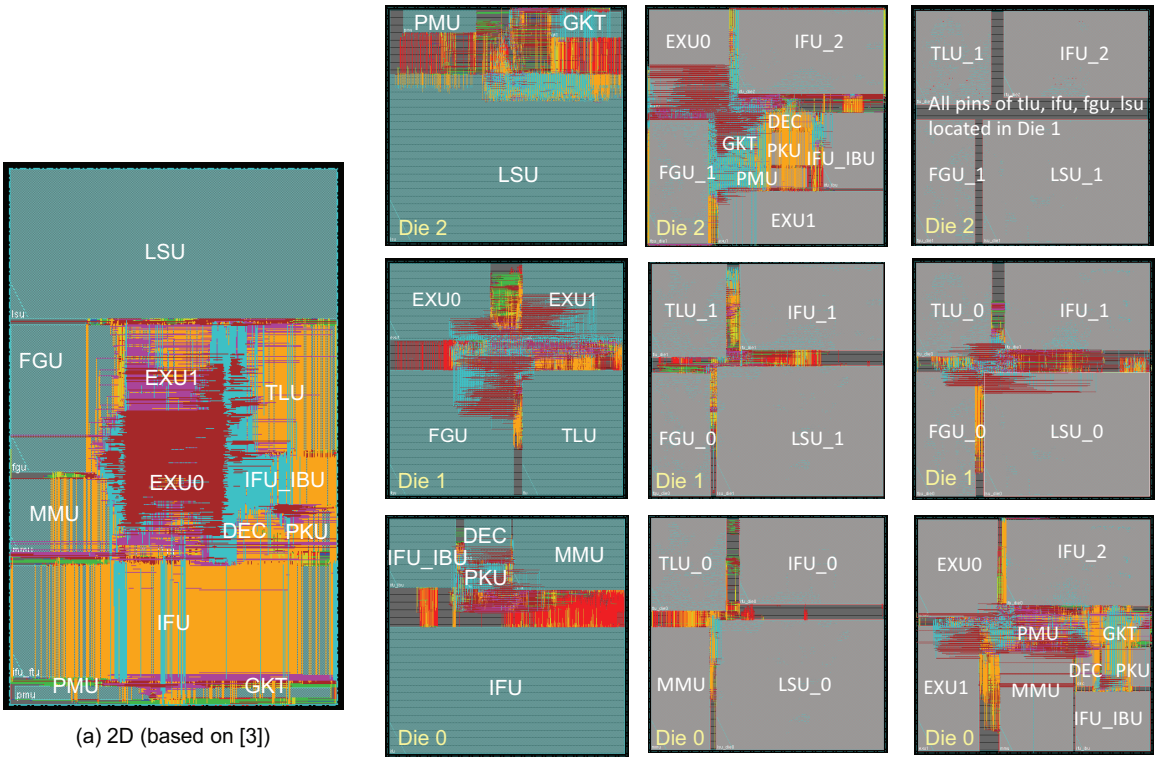
power reduction than F2B-only–block-folding design (top-level design quality: F2F+F2B V1 $>$ F2B-only–block-folding design $>$ F2F+F2B V2). Comparing the top-level design quality, V1 achieves -17.3% cell count and -20.4% wirelength reduction and -29.4% total top-level power reduction than F2B-only–block-folding design. Better top-level design quality leads to more power reduction in blocks, because it requires the blocks to use less resources to optimize the boundaries. Therefore the design quality impact by better top-level design cannot be ignored.

5.5.2.2 B2B+F2F Bonding on Folded Blocks

B2B+F2F bonding leads to a 3D layer using B2B bonding. Therefore, if top-level design uses F2F layer, blocks must use B2B layer for block-folding. Since Section 5.5.1.2 revealed the impact of B2B bonding on the top-level, it is important to study how the design quality of folded blocks change in B2B bonding. 2-tier standalone blocks were designed in T2 Core (LSU, FGU, TLU, and IFU), and results showed that F2B, B2B, and F2F bonding reduces block power compared to 2D (in average) by -5.9%, -2.4%, and -8.3%, respectively. B2B bonding shows the least power reduction among all other bonding styles. This is mainly due to the increased TSV RC parasitics (2x than F2B), occupying silicon area and TSV alignment issues in B2B bonding.

5.5.3 Overall Comparison

Table 26 compares all T2 core designs that have been done in this chapter based on whether block-folding technique is applied and the bonding style. GDSII layouts of our designs are illustrated in Figure 106, and designs that are not shown in the figure (such as non-folding–B2B+F2F) are based on a similar design as what is shown in Figure 106. First, a maximum of -36% power reduction is achieved in block-folded–F2F+F2B design. This is 14.8% more reduction than what was reported in [26], and the most power reduction reported in any previous studies. Second, block-folding provides more power reduction than non-folding. In terms of bonding style, F2F+F2B reduces most power, followed by



(a) 2D (based on [3])

(b) 3-tier non-folding (F2B-only bonding)

(c) 3-tier block-folding (F2B-only bonding)

(d) 3-tier block-folding (F2F+F2B bonding)

Figure 106: GDSII layouts of various 3-tier T2 Core designs: (a) 2D based on [26], (b) 3-tier non-folding in F2B-only, (c) 3-tier block-folding in F2B-only, and (d) 3-tier block-folding in F2F+F2B.

B2B+F2F and F2B-only style. However, to visualize more power reduction from these design techniques, more careful floorplanning and design must be done.

Table 26: Comparison among 3-tier T2 Core designs built with various options in folding and bonding styles. All folded designs target 4 blocks (LSU, IFU, TLU, and FGU) to be folded.

	2D	Non-Folding			Block-Folding		
	[26]	F2B-only	F2F+F2B	B2B+F2F	F2B-only	F2F+F2B	B2B+F2F
Clock period	1.5ns	1.5ns	1.5ns	1.5ns	1.5ns	1.5ns	1.5ns
Footprint (mm^2)	3.08	1.44 (-53.2%)	1.44 (-53.2%)	1.44 (-53.2%)	1.44 (-53.2%)	1.44 (-53.2%)	1.44 (-53.2%)
Si. Area (mm^2)	3.08	3.00 (-2.6%)	3.00 (-2.6%)	3.00 (-2.6%)	3.00 (-2.6%)	3.00 (-2.6%)	3.00 (-2.6%)
Wirelength (m)	22.4	14.3 (-36.2%)	13.8 (-38.4%)	14.1 (-37.1%)	13.4 (-40.2%)	13.0 (-42.0%)	13.2 (-41.1%)
# Cells	523.4K	403.9K (-22.8%)	394.3K (-24.7%)	395.7K (-24.4%)	370.9K (-29.1%)	368.8K (-29.5%)	370.4K (-29.2%)
# Buffers	221.7K	130.7K (-41.0%)	124.9K (-43.7%)	126.4K (-43.0%)	117.8K (-46.9%)	114.2K (-48.5%)	115.6K (-47.9%)
HVT cells	370.6K	377.4K	372.0K	374.4K	348.6K	346.0K	347.9K
	[70.7%]	[93.4%]	[94.3%]	[94.6%]	[93.9%]	[93.8%]	[93.9%]
# TSV	-	4,118	4,118	6,060	8,688	9,231	12,986
Total power (mW)	**348.3**	**248.1 (-28.8%)**	**242.6 (-30.3%)**	**244.4 (-29.8%)**	**229.7 (-34.0%)**	**223.1 (-36.0%)**	**227.4 (-34.7%)**
Cell power (mW)	71.6	62.6 (-12.6%)	62.1 (-13.3%)	62.0 (-13.4%)	54.1 (-24.4%)	54.0 (-24.6%)	54.1 (-24.4%)
Net power (mW)	175.7	117.3 (-33.2%)	113.3 (-35.5%)	115.0 (-34.5%)	107.7 (-38.7%)	102.0 (-41.9%)	105.4 (-40.0%)
Leak. power (mW)	101.1	68.2 (-32.5%)	67.1 (-33.6%)	67.4 (-33.3%)	67.9 (-32.8%)	67.2 (-33.5%)	67.9 (-32.8%)

5.6 Design Challenges in Full-Chip

This section describes the design challenges and results in full-chip 3-tier T2. Bigger design scale provides unique challenges in various metrics. For a thorough and comprehensive study, six different full-chip designs are provided based on block-folding and different bonding styles.

5.6.1 Full-chip OpenSPARC T2 Design

The full-chip scale OpenSPARC T2 consists of 53 blocks including eight SPARC cores (T2 Core), eight L2-cache data banks (L2D), eight L2-cache tags (L2T), eight L2-cache miss buffers (L2B), and a cache crossbar (CCX). Each block is synthesized with Synopsys 28nm cell libraries [74] as in T2 Core. Seven blocks that do not directly affect the CPU performance are removed from the implementation including five SerDes blocks, an electronic fuse, and a miscellaneous I/O unit. In addition, the PLL (analog block) is replaced in a clock control unit (CCU) by ideal clock sources. Thus, a total of 46 blocks are floorplanned. This study uses the same netlist as in the previous work [27], and the baseline 2D follows the full-chip T2 floorplan and designs done in [27]. However, since these designs did not have PDN, PDN is included in 2D and other designs and minor modifications are made to meet the timing.

5.6.2 Area Management Challenges

In IC designs, managing a small area is very important for low cost. Therefore, 3D ICs should also be designed in the smallest area possible. In section 5.3.2 and in previous studies [26], 3D ICs are reported to have the benefit of designing modules in a smaller area due to the reduced wirelength and buffer count. However, this statement may not always be true when designing ICs in full-chip scale. Table 27 shows how 3D ICs are bigger to their counterpart 2D in previous studies [44, 27, 54].

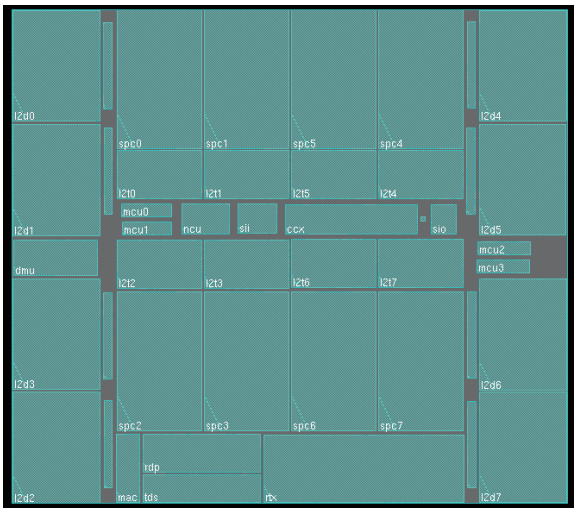
Notice that in full-chip scale studies, 3D ICs do not consume less silicon area than the

Table 27: Area comparison between 2D and 3D in full-chip level studies

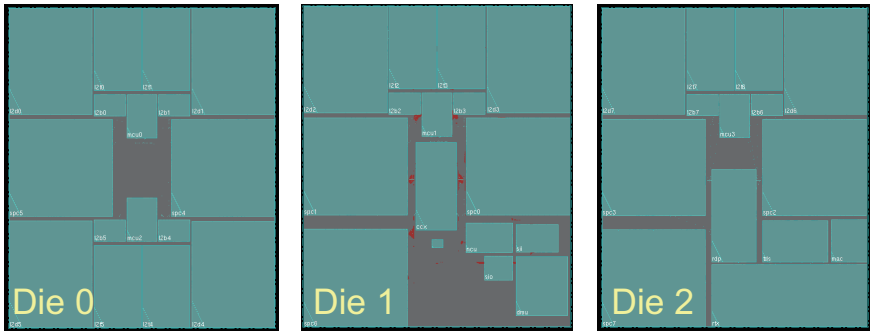
	[54]	[27]	[44]	This study
2D silicon area (mm^2)	5.5225	71.1	8.2	71.1
3D footprint (mm^2)	3.1725	38.4	4.1	24.3
3D silicon area (mm^2)	6.345	76.8	8.2	72.9
Increase rate (%)	14.9	8.0	0	2.5

2D. For example, in [54], 2D is 5.5225mm^2 and their 2-tier 3D is 6.345mm^2 (+14.9% more area). In the previous full-chip scale study done in T2 [27], 3D Uses 8.0% more silicon area than 2D. This is because of the following reason. This section will explain this in example of T2: Having 46 modules in full-chip level requires significant effort on floorplanning to maintain a small footprint. In T2, what is worse, the area difference between the biggest module (Core) and the smallest module (sio) is more than 16x. Therefore, managing a small-footprint floorplan is a challenging task in both 2D and 3D. However, floorplanning problem becomes more complicated in 3D ICs. For example, 2-tier 3D ICs require managing two seamless floorplans using only half of the number of total modules. Floorplanning becomes harder when there are less number of modules to place. In 3-tier 3D ICs, it becomes even more challenging because the designer must floorplan three surfaces using 1/3 of modules that the original 2D has. Many design constraints must be met in full-chip design, and these design constraints conflict with area management. However, note that a more complicated floorplanning problem in 3-tier do not always lead to more area consumption. In comparison with [27], 3-tier design in this study consumes less silicon area (72.9mm^2) than a 2-tier full-chip (76.8mm^2). Figure 107 shows a comparison between 2D and 3-tier full-chip floorplan. 3-tier full-chip consumes more silicon area (+2.5%), but note that the white space inside the 3-tier floorplan is also larger than 2D. In fact, all increased silicon area and the area saved from designing smaller modules in 3D remains as empty space since floorplanning in 3-tiers is a challenging task.

Having different chip sizes in different dies may be a viable solution to area management. While wafer-to-wafer (W2W, [24]) bonding cannot have different sized ICs on each



(a): 2D



(b): 3-tier 3D (F2F+F2B)

Figure 107: White space (= gray area) in T2 full-chip. (a) 2D floorplan (9mm x 7.9mm), (b) 3-tier 3D floorplan (4.5mm x 5.4mm). More silicon area used in 3D remains as white space due to floorplanning challenges.

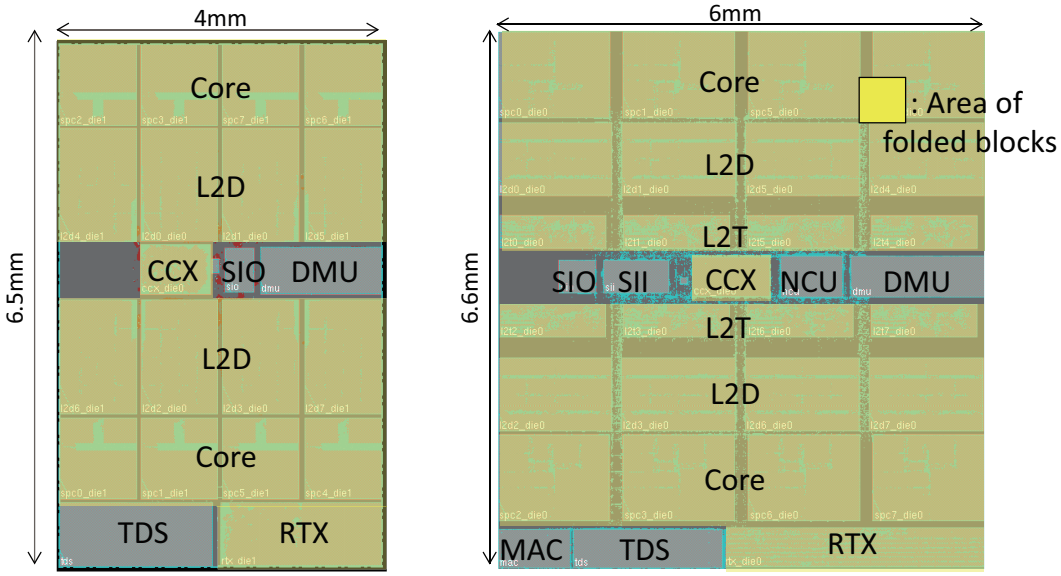
tier, chip-to-wafer (C2W, [46]) or chip-to-chip (C2C, [67]) bonding provides possibilities to use differently-sized dies in different tiers. However, C2W and C2C bonding comes with inferior accuracy and cost than using W2W bonding. Smaller dies are required to be handled with more advanced equipments, and in addition to this, handling smaller chip-scale dies result in reduced placement accuracy [25]. In some cases, smaller dies may not be able to be bonded in C2C or C2W style due to the equipments. Therefore, designers must choose the 3D partition and floorplan wisely based on various design factors including these different chip bonding styles.

5.6.3 Block-Folding in Full-Chip

Block-folding in 3-tier becomes more challenging in full-chip due to the bigger design complexity. This section reports how block-folding is different from 2-tier and describes the proposed block-folding techniques.

5.6.3.1 How Many Blocks Can We Fold?

In addition to regarding area balance in Section 5.3.1, the actual area that can be used for folding reduces due to the reduced footprint. Therefore, designers must properly choose what blocks to fold based on power reduction and floorplanning benefits. Figure 108 shows how the area for folding reduces in 3-tier full-chip layout. As in (b), 2-tier 3D allows to fold five different modules (Core, RTX, L2D, L2T, CCX) [27]. Because of the reduced footprint in the folding die in Die 1, 3-tier only allows to fold four modules. However, notice that different number of tiers stem distinctive challenges. For example, a 4-tier 3D will have different folding constraints of a 3-tier design. E.g., 4-tier design can use Die0-Die1 and Die2-Die3 for folding since this would not overlap to each other.



(a) 3-tier full-chip (4 blocks folded: Core, CCX, L2D, RTX)

(b) 2-tier full-chip (5 blocks folded: Core, CCX, L2D, RTX, L2T)

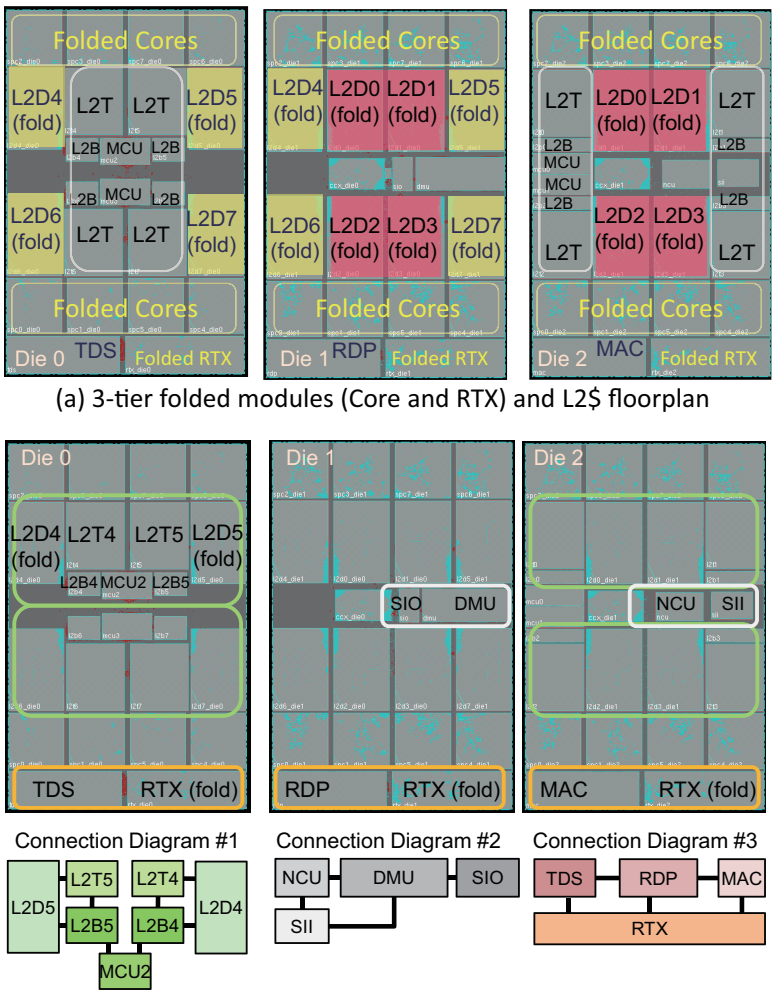
Figure 108: How folding area reduces in 3-tier designs. Footprint reduction in 3-tier leads to less folded blocks. (a) Die 1 in 3-tier, (b) Die 1 in 2-tier [27].

5.6.3.2 Block-Folding Design Strategies in Full-Chip

This section describes how 3-tier full-chip floorplan with block-folding is done considering all challenges described previous sections. Though this is an example for OpenSPARC T2 architecture, the basic ideas can further extend to other microprocessor architectures as well. Figure 109 and 110 shows how the proposed block-folding design strategy is applied in the layout. First, 3-tier-folding is done only on Cores and RTX. 3-tier folding may provide more power reduction than 2-tier folding. However, a 3-tier folded block becomes a floorplan/routing blockage in all 3-tiers. These folded blocks cause routing problems when they are placed in the middle of the die. Thus, these 3-tier blocks are placed on the top and bottom of the floorplan.

Second, CCX and L2Ds are folded in 2-tiers. L2Ds do not provide an impressive power reduction when it is folded, but it is folded for a better top-level floorplan. When deciding a floorplan, having huge-sized modules is not preferable because of the reduced design freedom it provides on the top-level. Especially for hard modules that the designer cannot change its size freely, it is more advantageous to have its size as small as possible. L2D is a module that consists of 32 memory macros so that the size changing is not easy. Therefore, L2Ds are folded into 2-tiers. L2Ds were the biggest module inside the top-level block-folding floorplan before folding, but the size of its 2-tier footprint is now comparable to other modules in the top-level floorplan.

Third, modules that are heavily connected to each other are gathered together. In fact, L2$s (L2D, L2T, L2B, and MCU) are heavily connected to each other. To utilize the block-folding space efficiently, Die 1 is used for folded L2Ds, and other L2$s are placed on Die 0 and Die 2. However, folding restriction from Die 1 limits some L2Ds being placed on its sub-optimal locations. Therefore, Die0-die1 L2Ds are chosen to be placed on the side which provides the best floorplan for L2$s, and Die1-Die2 L2Ds are placed on the middle of the chip. However, due to this, the L2$ floorplan in Die2 becomes inferior than Die0. For best L2$ connections, L2D4 - L2D7 I/Os are assigned on Die 0 and L2D0 - L2D3 I/Os



(a) 3-tier folded modules (Core and RTX) and L2$ floorplan

(b) Highly-connective modules are placed close to each other

Figure 109: Full-chip block-folding floorplan strategies: (a) 3-tier folded modules and L2$ floorplan. Die 1 is utilized to place folded L2Ds, and other L2$s are placed on Die 0 and Die 2. Corresponding L2D pins are placed on each dies. (b) How highly-connective modules are placed closely to each other and its connection diagram. (c) L2T-CCX and CCX-Core I/O pin assignment to reduce congestion.

are assigned on Die 2. In addition to L2$s, NIU modules (TDS, RDP, MAC, and RTX) are heavily connected to each other and do not have many connections to other modules. Therefore, all NIU modules are gathered on the bottom of the chip. DMU, NCU, and SIU modules (SIO and SII) have many connections to each other, so they are gathered as well. Finally, I/O pins of the folded modules are properly managed. In the OpenSPARC architecture, Cores do not directly connect with L2$s. In fact, most of the Core I/Os connect to CCX, and CCX connects to L2Ts. Having this architecture, and knowing that L2Ts are

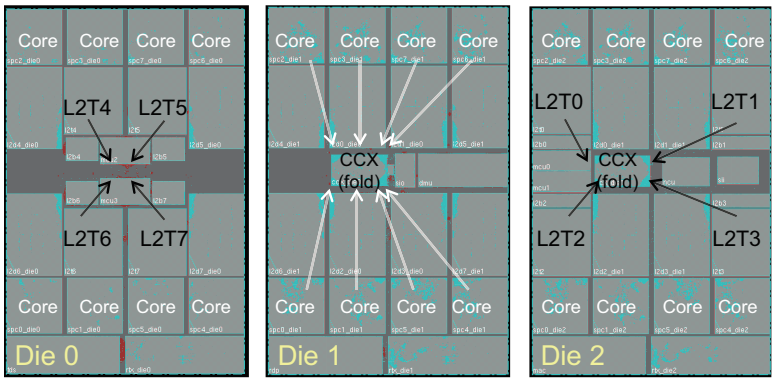


Figure 110: Full-chip block-folding floorplan strategy: L2T-CCX and CCX-Core I/O pin assignment to reduce congestion.

placed on Die 0 and Die 2, Core I/Os that connect to CCX must be managed properly. By placing Core I/Os on Die 1 and placing CCX I/Os that connect to L2Ts and Cores on the same die of its connecting module, significant congestion between CCX-L2T and CCX-Core can be resolved in top-level design.

5.6.4 Managing Bonding Styles in Full-Chip

Managing an adequate bonding style is also important for more power reduction in full-chip designs. Comparing Table 26 and Table 28, some differences are noticed that occur in non-folded full-chip designs compared to single core designs: First, F2F+F2B bonding do not provide significant power reduction over F2B-only bonding. Second, the power penalty from F2F+F2B to B2B+F2F is not significant.

5.6.4.1 Advantages of F2F Bonding

In non-folded T2 Core, -1.5% more power reduction was achieved when F2F+F2B bonding was chosen over F2B-only (Table 26). However, in non-folded T2 full-chip, only -0.6% is obtained more. This is explained through the following: In core, top-level routing required many I/Os to be connected between modules. Due to this, non-folded Core must have TSVs in particular spots. Therefore, TSVs were crowded on its sub-optimal locations (see Figure 103). However, in the full-chip, I/Os that are connecting to other blocks are relatively

sparse compared to Core due to careful I/O managing. Note that TSV count in Die 0 is 2176 in Core and 2356 in full-chip. Despite that the design size increased by more than 20x, TSV count is similar to each other.

To obtain more power reduction from F2F bonding, the initial F2B design requires to (1) have many TSVs and (2) these TSVs should be congested so that it cannot find its optimal locations. F2F+F2B Core could benefit more from F2F bonding since it met these two criteria. However, I/Os are managed to have less TSVs with less congestion in the full-chip. In addition, full-chip design has signficant white space for TSVs. TSVs already find its optimal spot during TSV placement. Therefore, significant benefit is not shown from F2F bonding. Comparing Figure 103 from Figure 111, notice that TSVs in full-chip are already placed in its optimal location. In summary, due to the good TSV locations full-chip F2B-only non-folded design provide, it does not show significant power reduction when full-chip design moves to F2F+F2B bonding.

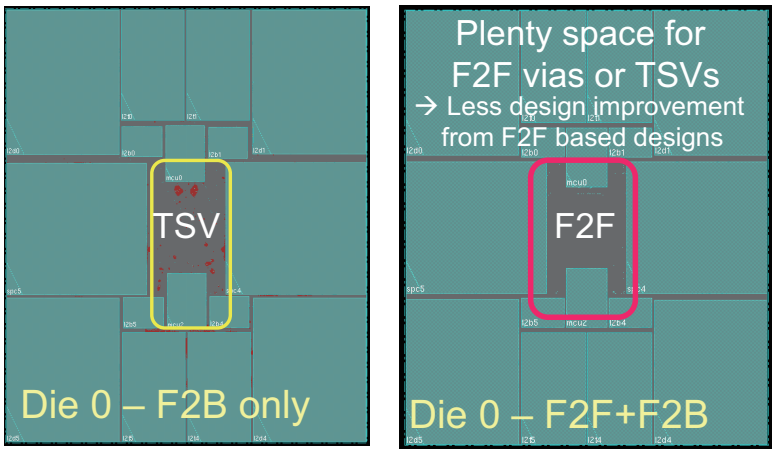


Figure 111: TSV/F2F placement in full-chip. Because TSVs are placed in its optimal locations (left) due to less congestion and large whitespace, F2F bonding (right) do not provide significant benefits over TSVs.

5.6.4.2 Managing B2B Bonding

In non-folded T2 Core, B2B+F2F bonding consumes +0.5% more power than F2F+F2B bonding. However, in non-folded full-chip, B2B+F2F bonding consumes only +0.1% more

power than F2F+F2B bonding. This is because B2B+F2F design did not have many issues with placing TSVs on both dies. B2B bonding becomes a significant design issue when TSVs cannot find white spaces to be placed on both dies. However, in full-chip level where TSVs have sufficient space to be placed, B2B bonding will not become a significant handicap compared to F2B bonding style. Notice that in the full-chip design in this study, TSVs can easily be placed on both sides of chip, and this leads to almost negligible penalty when using B2B bonding. In summary, block-level full-chip designs did not show significant difference between different bonding styles. Maximum bonding style impacts came from block-folded full-chip designs, and this is because of the design benefits/issues that rise from more 3D connections.

5.6.5 Overall Comparison in Full-Chip

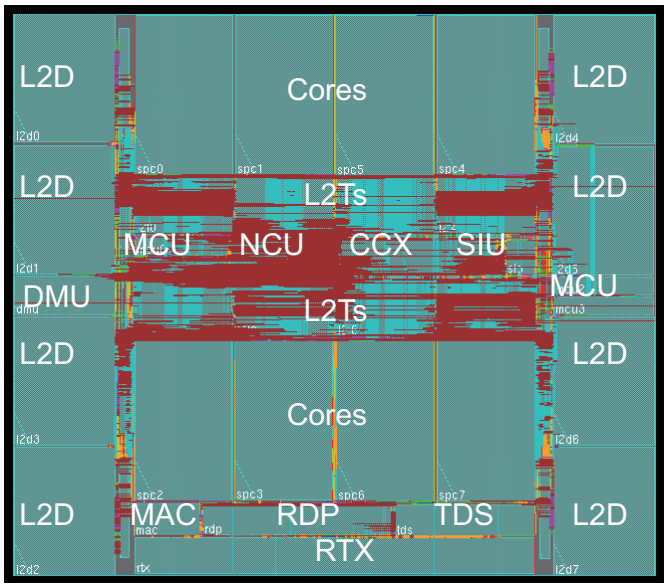
Table 28 compares all full-chip designs that have been done based on whether block-folding technique is applied and the bonding style. GDSII layouts of the designs done in this study are illustrated in Figure 112, and designs that are not shown in the figure (F2B-only and B2B+F2F bonding styles on both non-folded and block-folded full-chip) are based on a similar floorplan of what is shown in Figure 112. First, this study emphasizes that a maximum of -27.2% power reduction has been achieved in block-folded-F2F+F2B design. This is -6.9% more reduction than what was reported in [27]. Note that the power reduction from 3-tier design is almost similar to one technology node difference. This study also emphasizes that this is the maximum power reduction reported in any kind of full-chip studies. Second, similar as T2 Core results in Section 5.5.3, block-folding provides more power reduction than non-folding. In terms of bonding style, F2F+F2B reduces most power, followed by B2B+F2F and F2B-only style. For maximum power reduction in 3-tier 3D ICs, all 3D design techniques we have mentioned in this paper such as floorplanning, pin assignment, block-folding, and TSV assignment should be carefully managed.

Table 28: Full-chip comparison among 3-tier 3D IC designs built with various options in folding and bonding styles. All folded designs target 4 blocks (Core, RTX, CCX, L2D) to be folded.

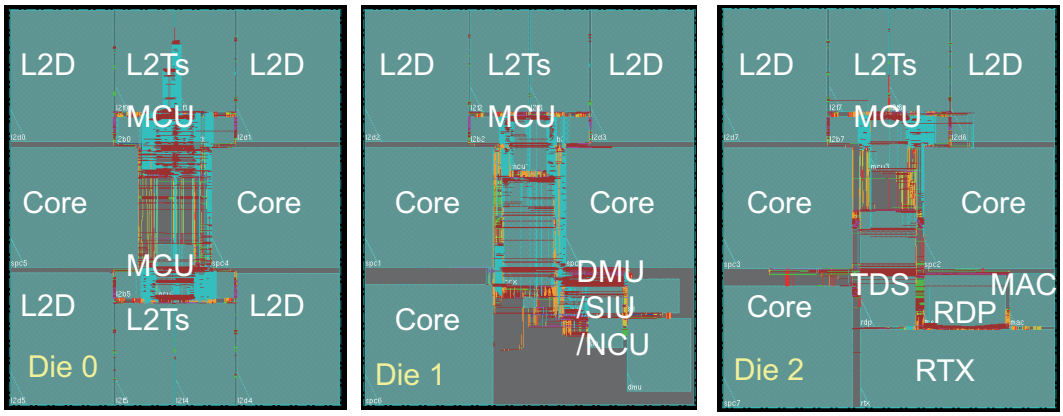
	2D	Non-Folding			Block-Folding		
	[26]	F2B-only	F2F+F2B	B2B+F2F	F2B-only	F2F+F2B	B2B+F2F
Clock period	2ns	2ns	2ns	2ns	2ns	2ns	2ns
Footprint (mm^2)	71.1	24.3 (-65.8%)	24.3 (-65.8%)	24.3 (-65.8%)	25.8 (-63.7%)	25.8 (-63.7%)	25.8 (-63.7%)
Si. Area (mm^2)	71.1	72.9 (+2.5%)	72.9 (+2.5%)	72.9 (+2.5%)	77.4 (+8.9%)	77.4 (+8.9%)	77.4 (+8.9%)
Wirelength (m)	343.0	248.1 (-27.7%)	247.0 (-28.0%)	246.6 (-28.1%)	234.4 (-31.7%)	227.1 (-33.8%)	228.7 (-33.3%)
# Cells	7.56M	6.48M (-14.3%)	6.43M (-14.9%)	6.44M (-14.8%)	5.99M (-20.8%)	5.92M (-21.6%)	5.95M (-21.3%)
# Buffers	3.05M	1.97M (-35.4%)	1.92M (-37.0%)	1.93M (-36.7%)	1.69M (-44.6%)	1.62M (-46.8%)	1.65M (-45.9%)
HVT cells	6.57M	6.06M	6.02M	6.03M	5.46M	5.44M	5.44M
	[86.9%]	[93.5%]	[92.9%]	[93.6%]	[91.1%]	[91.8%]	[91.4%]
# TSV	-	4,599	4,599	6,842	55,142	82,743	93,185
Total power (W)	**8.614**	**6.695 (-22.3%)**	**6.649 (-22.8%)**	**6.654 (-22.7%)**	**6.406 (-25.6%)**	**6.275 (-27.2%)**	**6.335 (-26.5%)**
Cell power (W)	1.757	1.525 (-13.2%)	1.521 (-13.4%)	1.523 (-13.1%)	1.431 (-18.6%)	1.421 (-19.1%)	1.425 (-18.9%)
Net power (W)	4.770	3.327 (-30.3%)	3.294 (-30.9%)	3.290 (-31.0%)	3.231 (-32.3%)	3.120 (-34.6%)	3.167 (-33.6%)
Leak. power (W)	2.087	1.843 (-11.7%)	1.835 (-12.1%)	1.841 (-11.8%)	1.744 (-16.4%)	1.734 (-16.9%)	1.743 (-16.5%)

5.7 Summary

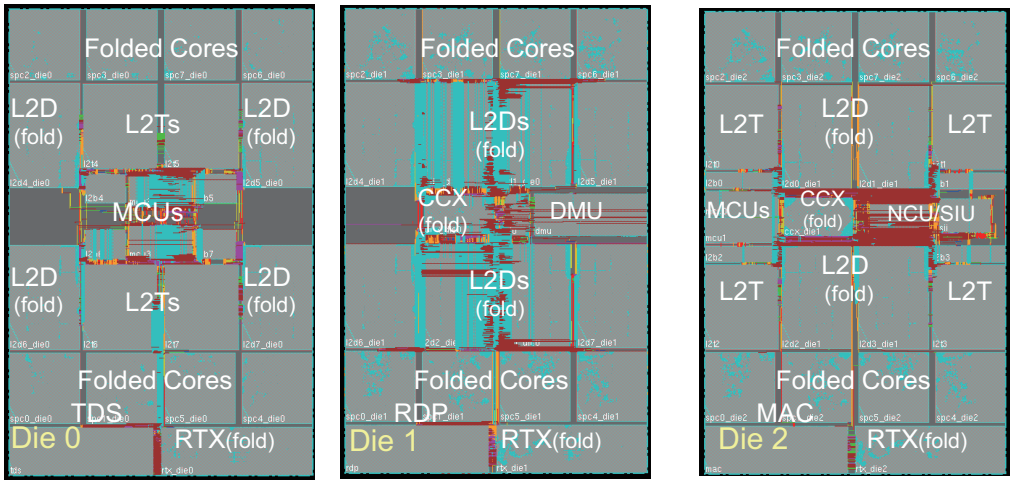
This chapter demonstrated power reduction benefits that 3-tier 3D IC design provides in OpenSPARC T2. First, it was shown that one additional tier in 3-tier 3D ICs offers more power savings than 2-tier 3D ICs. Second, 3-tiers can be bonded in various mixed styles, and these various styles provide additional power reduction. However, more careful floor-planning, TSV management, and block-folding considerations are required. Lastly, to demonstrate the maximum power reduction of 3-tier 3D ICs, this study developed CAD tools that seamlessly integrate into commercial 2D tools for design and optimization. With aforementioned methods and design techniques combined, this study has achieved -36.0% total power saving against the 2D counterpart in T2 Core, and -27.2% total power saving in full-chip T2 microprocessor.



(a) 2D



(b) Non-fold



(c) Block-fold

Figure 112: GDSII layouts of various full-chip 3-tier 3D IC designs in F2F+F2B bonding: (a) 2D based on [27], (b) 3-tier non-folding, and (c) 3-tier block-folding.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

Three-dimensional integrated circuits (3D ICs) have gained significant attention over the past decade as a technology that can facilitate the continuation of the advances guided by the Moore's law. Through many studies, including this work, 3D ICs are expected to providing more computing capability in low power, more data transfer bandwidth, heterogeneous integration, and so on. Since previous studies have not taken system level components such as silicon interposers/packages and PCBs into account, this work has developed many co-design methodologies that could provide more reliable analysis in the system level. This dissertation presented the following studies:

- A design methodology of co-simulating IR-drop noise for 3D IC, silicon interposer, and PCB simultaneously.
- A thermal analysis methodology for analog/digital mixed signal systems.
- TSV-to-TSV coupling and its impact on ICs in comparison with package/PCB elements.
- Design methodologies and algorithms for full-chip TSV-to-TSV coupling analysis.
- Face-to-face parasitic analysis and design methodologies for full-chip extraction.
- Design methodologies and CAD tools for 3-tier 3D ICs and its power reduction.

The co-IR-drop noise analysis provided a holistic platform to analyze IR-drop in a system level including silicon interposer/package and PCB simultaneously. From the proposed analysis platform, significant optimization and turn-around time between different domains (IC, package, PCB) could be saved. However, in addition to IR-drop noise, the AC droop

noise is also a significant problem in PDNs. Therefore, for an accurate PDN analysis, a holistic model including inductance and decoupling capacitors for AC droop noise must also be studied.

The thermal analysis methodology in this dissertation for the first time co-analyzed temperature in analog/digital mixed signal systems including silicon interposers in one platform. It demonstrated how significant heat it generates in integrated voltage regulators, and provided some techniques to reduce the temperature when the system is including voltage regulators. However, the performance reduction by adjusting the floorplan of the system components were not analyzed. To validate the actual impact of the design techniques proposed in this work, performance reduction from the floorplanning must be considered.

The TSV-to-TSV coupling study in this dissertation provided an accurate analysis of the actual coupling behavior inside ICs when there are more than hundreds of TSVs. It showed how TSV coupling in ICs and package/PCBs are different. Then, it provided an algorithm and a methodology of analyzing coupling between multiple TSVs in full-chip scale. However, since the silicon substrate is getting thinner, more coupling from the device-to-TSV would occur. For a comprehensive coupling study, this must be considered.

Face-to-face parasitic analysis in this dissertation studied the impact of 3D capacitances that were not existing in 2D ICs. It provided design guidelines and full-chip level analyses of how far the dies should be in order to see minimum impact from face-to-face bonding. However, in the industry's perspective, it may not be possible to have a detailed IC information on two dies when the vendor is different. Therefore, an approach to accurately extract parasitics even with less IC information should be developed.

3-Tier study in this dissertation provided design methodologies and CAD tools for 3-tier block-level 3D ICs and showed significant power reduction. Based on careful floorplanning, block-folding, pin assignment, and various bonding styles, significant power reduction can be achieved. However, 3-tier 3D ICs expect to have thermal issues. Thus, thermal study for many-tier 3D ICs must be followed. In addition, circuit techniques developed

for power reduction have not been applied in 3D ICs. It is expected to reduce more power by these techniques, and novel 3D IC power reduction techniques should be proposed for more power reduction.

PUBLICATIONS

This dissertation is based on and/or related to the work and results presented in the following publications in print:

[1] **Taigon Song**, Chang Liu, Dae Hyun Kim, Jonghyun Cho, Joohee Kim, Jun So Pak, Seungyoung Ahn, Joungho Kim, Kihyun Yoon, and Sung Kyu Lim, "Analysis of TSV-to-TSV Coupling with High-Impedance Termination in 3D ICs," in *Proceedings IEEE International Symposium on Quality Electronic Design*, pp. 1–7, March 2011.

[2] Chang Liu, **Taigon Song**, and Sung Kyu Lim, "Signal Integrity Analysis and Optimization for 3D ICs," in *IEEE International Symposium on Quality Electronic Design*, pp. 1–7, March 2011.

[3] Chang Liu, **Taigon Song**, Jonghyun Cho, Joohee Kim, Joungho Kim, and Sung Kyu Lim, "Full-Chip TSV-to-TSV Coupling Analysis and Optimization in 3D IC," in *ACM/IEEE Design Automation Conference*, pp. 783–788, June 2011.

[4] **Taigon Song** and Sung Kyu Lim, "Co-design and Co-simulation of 3D IC and Silicon Interposer Power Distribution Network," in *IEEE Workshop on Chip-Packaging Co-Design for High Performance Electronic Systems*, September 2011.

[5] **Taigon Song** and Sung Kyu Lim, "A Fine-Grained Co-Simulation Methodology for IR-drop Noise in Silicon Interposer and TSV-based 3D IC," in *IEEE Electrical Performance of Electronic Packaging and Systems*, pp. 239–242, October 2011.

[6] **Taigon Song**, Noah Sturcken, Krit Athikulwongse, Kenneth Shepard, and Sung Kyu Lim, "Thermal Analysis and Optimization of 2.5-D Integrated Voltage Regulator,"

in *IEEE Electrical Performance of Electronic Packaging and Systems*, pp. 25–28, October 2012.

[7] **Taigon Song**, Chang Liu, Yarui Peng, and Sung Kyu Lim, "Full-Chip Multiple TSV-to-TSV Coupling Extraction and Optimization in 3D ICs," in *ACM/IEEE Design Automation Conference*, pp. 1–7, June 2013.

[8] **Taigon Song**, Chang Liu, Yarui Peng, and Sung Kyu Lim, "Multiple TSV-to-TSV Coupling Extraction, Analysis, and Optimization in 3D ICs," in *SRC TECHCON Conference*, September 2013.

[9] Moongon Jung, Young-Joon Lee, **Taigon Song**, Yang Wan, and Sung Kyu Lim, "Design Methodologies for Low Power 3D Processors," in *SRC TECHCON Conference*, September 2013.

[10] Moongon Jung, **Taigon Song**, Yang Wan, Young-Joon Lee, Debabrata Mohapatra, Hong Wang, Greg Taylor, Devang Jariwala, Vijay Pitchumani, Patrick Morrow, Clair Webb, Paul Fischer, and Sung Kyu Lim, "How to Reduce Power in 3D IC Designs: A Case Study with OpenSPARC T2 Core," in *IEEE Custom Integrated Circuits Conference*, pp. 1–4, September 2013.

[11] Yarui Peng, **Taigon Song**, Dusan Petranovic and Sung Kyu Lim, "On Accurate Full-Chip Extraction and Optimization of TSV-to-TSV Coupling Elements in 3D ICs," in *IEEE International Conference on Computer-Aided Design*, pp. 281–288, November 2013.

[12] Moongon Jung,**Taigon Song**, Yang Wan, Yarui Peng, and Sung Kyu Lim, "On Enhancing Power Benefits in 3D ICs: Block Folding and Bonding Styles Perspective," in *ACM/IEEE Design Automation Conference*, pp.1–6, June 2014.

[13] **Taigon Song** and Sung Kyu Lim, "Die-to-Die Parasitic Extraction Targeting Face-to-Face Bonded 3D ICs," in *Journal of Information and Communication Convergence*

Engineering, – to appear.

[14] **Taigon Song**, Arthur Nieuwoudt, Yun Seop Yoo, and Sung Kyu Lim, "Impact of Irregular Geometries on Low-k Dielectric Breakdown," in *IEEE Electronic Components and Technology Conference*, pp.529–536, ,May 2015.

[15] Yarui Peng, Moongon Jung, **Taigon Song**, Yang Wan, and Sung Kyu Lim, "Thermal Impact Study of Block Folding and Face-to-Face Bonding in 3D IC," in *IEEE International Interconnect Technology Conference*, May 2015.

[16] **Taigon Song**, Moongon Jung, Yang Wan, Yarui Peng, and Sung Kyu Lim, "3D IC Power Benefit Study Under Practical Design Considerations," in *IEEE International Interconnect Technology Conference*, May 2015.

[17] **Taigon Song**, Shreepad Panth, Yoo-Jin Chae, and Sung Kyu Lim, "Three-Tier 3D ICs for More Power Reduction: Strategies in CAD, Design, and Bonding Selection," in *IEEE/ACM International Conference on Computer-Aided Design*, – to appear.

[18] Yarui Peng, **Taigon Song**, Dusan Petranovic, and Sung Kyu Lim, "Full-chip Inter-die Parasitic Extraction in Face-to-Face-Bonded 3D ICs, " in *IEEE/ACM International Conference on Computer-Aided Design*, – to appear.

[19] Moongon Jung, **Taigon Song**, Yarui Peng, and Sung Kyu Lim, "Fine-Grained 3D IC Partitioning Study with A Multi-core Processor," in *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, – to appear.

[20] **Taigon Song**, Chang Liu, Yarui Peng, and Sung Kyu Lim, "Impact of Through-Silicon-Via Scaling on the Wirelength Distribution of Current and Future 3D ICs," in *IEEE Transactions on Very Large Scale Integration Systems*, – to appear.

In addition, the author has completed studies unrelated to this dissertation presented in the following publications in print:

[1] Dae Hyun Kim, Krit Athikulwongse, Michael B. Healy, Mohammad M. Hossain, Moongon Jung, Ilya Khorosh, Gokul Kumar, Young-Joon Lee, Dean L. Lewis, Tzu-Wei Lin, Chang Liu, Shreepad Panth, Mohit Pathak, Minzhen Ren, Guanhao Shen, **Taigon Song**, Dong Hyuk Woo, Xin Zhao, Joungho Kim, Ho Choi, Gabriel H. Loh, Hsien-Hsin S. Lee, and Sung Kyu Lim, "3D-MAPS: 3D Massively Parallel Processor with Stacked Memory," in *IEEE International Solid-State Circuits Conference*, February 2012.

[2] Darryl Kostka, **Taigon Song**, and Sung Kyu Lim, "3D IC-Package-Board Co-analysis Using 3D EM Simulation for Mobile Applications," in *IEEE Electronic Components and Technology Conference*, pp. 2113–2120, May 2013.

[3] Yarui Peng, **Taigon Song**, Dusan Petranovic, and Sung Kyu Lim, "Silicon Effect-aware Full-chip Extraction and Mitigation of TSV-to-TSV Coupling," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 33, No. 12, pp. 1900–1913, 2014.

[4] Dae Hyun Kim, Krit Athikulwongse, Michael B. Healy, Mohammad M. Hossain, Moongon Jung, Ilya Khorosh, Gokul Kumar, Young-Joon Lee, Dean L. Lewis, Tzu-Wei Lin, Chang Liu, Shreepad Panth, Mohit Pathak, Minzhen Ren, Guanhao Shen, **Taigon Song**, Dong Hyuk Woo, Xin Zhao, Joungho Kim, Ho Choi, Gabriel H. Loh, Hsien-Hsin S. Lee, and Sung Kyu Lim, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," in *IEEE Transactions on Computers*, Vol. 64, No. 1, pp. 112–125, 2015.

[5] **Taigon Song** and Sung Kyu Lim, "Full-chip Power Performance Benefit Study for Carbon Nanotube-based Circuits," in *Journal of Information and Communication Convergence Engineering*, – to appear.

REFERENCES

[1] "3D IC & TSV Report : Cost, Technologies & Markets, Yole Development," 2007.

[2] "3D IC & TSV interconnects 2012 Business update, Yole Dvelopment, Semicon Taiwan 2012," 2012.

[3] ABEDINPOUR, S., BAKKALOGLU, B., and KIAEI, S., "A Multi-Stage Interleaved Synchronous Buck Converter with Integrated Output Filter in a 0.18/spl mu/ SiGe process," in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pp. 1398–1407, Feb 2006.

[4] AHMED, K. and SCHUEGRAF, K., "Transistor wars," *Spectrum, IEEE*, vol. 48, pp. 50–66, November 2011.

[5] ALLEC, N., HASSAN, Z., SHANG, L., DICK, R., and YANG, R., "ThermalScope: Multi-scale thermal analysis for nanometer-scale integrated circuits," in *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, pp. 603–610, Nov 2008.

[6] AUTH, C., ALLEN, C., BLATTNER, A., BERGSTROM, D., BRAZIER, M., BOST, M., BUEHLER, M., CHIKARMANE, V., GHANI, T., GLASSMAN, T., GROVER, R., HAN, W., HANKEN, D., HATTENDORF, M., HENTGES, P., HEUSSNER, R., HICKS, J., INGERLY, D., JAIN, P., JALOVIAR, S., JAMES, R., JONES, D., JOPLING, J., JOSHI, S., KENYON, C., LIU, H., MCFADDEN, R., MCINTYRE, B., NEIRYNCK, J., PARKER, C., PIPES, L., POST, I., PRADHAN, S., PRINCE, M., RAMEY, S., REYNOLDS, T., ROESLER, J., SANDFORD, J., SEIPLE, J., SMITH, P., THOMAS, C., TOWNER, D., TROEGER, T., WEBER, C., YASHAR, P., ZAWADZKI, K., and MISTRY, K., "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *VLSI Technology (VLSIT), 2012 Symposium on*, pp. 131–132, June 2012.

[7] BLACK, B., ANNAVARAM, M., BREKELBAUM, N., DEVALE, J., JIANG, L., LOH, G., MCCAULEY, D., MORROW, P., NELSON, D., PANTUSO, D., REED, P., RUPLEY, J., SHANKAR, S., SHEN, J., and WEBB, C., "Die Stacking (3D) Microarchitecture," in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, pp. 469–479, Dec 2006.

[8] CADIX, L., ROUSSEAU, M., FUCHS, C., LEDUC, P., THUAIRE, A., EL FARHANE, R., CHAABOUNI, H., ANCIANT, R., HUGUENIN, J.-L., COUDRAIN, P., FARCY, A., BERMOND, C., SILLON, N., FLECHET, B., and ANCEY, P., "Integration and frequency dependent electrical modeling of through silicon vias (tsv) for high density 3dics," in *Interconnect Technology Conference (IITC), 2010 International*, pp. 1–3, June 2010.

[9] CHANG, Y.-J., CHUANG, H.-H., LU, Y.-C., CHIOU, Y.-P., WU, T.-L., CHEN, P.-S., WU, S.-H., KUO, T.-Y., ZHAN, C.-J., and LO, W.-C., "Novel crosstalk modeling for multiple through-silicon-vias (TSV) on 3-D IC: Experimental validation and application to Faraday cage design," in *Electrical Performance of Electronic Packaging and Systems (EPEPS), 2012 IEEE 21st Conference on*, pp. 232–235, Oct 2012.

[10] CHENG, D. K., *Field and Wave Eletromagnetics*. Boston, MA: Addison Wesley, second ed., 1992.

[11] CHO, J., SHIM, J., SONG, E., PAK, J. S., LEE, J., LEE, H., PARK, K., and KIM, J., "Active circuit to through silicon via (TSV) noise coupling," in *Electrical Performance of Electronic Packaging and Systems, 2009. EPEPS '09. IEEE 18th Conference on*, pp. 97–100, Oct 2009.

[12] CHUANG, H.-H., WU, T.-L., HONG, M.-Z., HSU, D., HUANG, R., HSIAO, L. C., and WU, T.-L., "Power integrity chip-package-PCB co-simulation for I/O interface of DDR3 high-speed memory," in *Advanced Packaging and Systems Symposium, 2008. EDAPS 2008. Electrical Design of*, pp. 31–34, Dec 2008.

[13] CHUANG, H.-C., LI, H.-F., LIN, Y.-S., LIN, Y.-H., and HUANG, C.-S., "The development of an atom chip with through silicon vias for an ultra-high-vacuum cell," *Journal of Micromechanics and Microengineering*, vol. 23, no. 8, p. 085004, 2013.

[14] CURRAN, B., NDIP, I., GUTTOVSKI, S., and REICHL, H., "The impacts of dimensions and return current path geometry on coupling in single ended Through Silicon Vias," in *Electronic Components and Technology Conference, 2009. ECTC 2009. 59th*, pp. 1092–1097, May 2009.

[15] DISCO, C. and VAN DER MEULEN, B., *Getting new technologies together*. New York: Wlater de Gruyter, 1969.

[16] DORSEY, P., "Xilinx Stacked Silicon Interconnect Technology Delivers Breakthrough FPGA Capacity, Bandwidth, and Power Efficiency," 2010.

[17] GHAIDA, R., TORRES, G., and GUPTA, P., "Single-mask double-patterning lithography for reduced cost and improved overlay control," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 24, pp. 93–103, Feb 2011.

[18] HA, M., SRINIVASAN, K., and SWAMINATHAN, M., "Chip-package co-simulation with multiscale structures," in *Electrical Performance of Electronic Packaging, 2008 IEEE-EPEP*, pp. 339–342, Oct 2008.

[19] HALDER, S., TRUFFERT, V., VAN DEN HEUVEL, D., LERAY, P., CHENG, S., MCINTYRE, G., SAH, K., BROWN, J., PARISI, P., and POLLI, M., "Inspection challenges for triple patterning at sub-14 nm nodes with broadband plasma inspection platforms," in *Advanced Semiconductor Manufacturing Conference (ASMC), 2015 26th Annual SEMI*, pp. 19–22, May 2015.

[20] HAOND, M., "20 nm fdsoi process and design platforms for high performance/ low power systems on chip," in *SOI Conference (SOI), 2012 IEEE International*, pp. 1–2, Oct 2012.

[21] HAZUCHA, P., SCHROM, G., HAHN, J., BLOECHEL, B., HACK, P., DERMER, G., NARENDRA, S., GARDNER, D., KARNIK, T., DE, V., and BORKAR, S., "A 233-MHz 80efficient four-phase DC-DC converter utilizing air-core inductors on package," *Solid-State Circuits, IEEE Journal of*, vol. 40, pp. 838–845, April 2005.

[22] HELLEMANS, A., "Ring around the nanowire [News]," *Spectrum, IEEE*, vol. 50, pp. 14–16, May 2013.

[23] HSIEH, M.-C., YU, C.-K., and WU, S.-T., "Thermo-mechanical simulative study for 3D vertical stacked IC packages with spacer structures," in *Semiconductor Thermal Measurement and Management Symposium, 2010. SEMI-THERM 2010. 26th Annual IEEE*, pp. 47–54, Feb 2010.

[24] HUYGHEBAERT, C., VAN OLMEN, J., CIVALE, Y., PHOMMAHAXAY, A., JOURDAIN, A., SOOD, S., FARRENS, S., and SOUSSAN, P., "Cu to Cu interconnect using 3D-TSV and wafer to wafer thermocompression bonding," in *Interconnect Technology Conference (IITC), 2010 International*, pp. 1–3, June 2010.

[25] J. YANNOU ET AL, "SET IS WELL POSITIONED AND PREPARED TO ADDRESS THE CHALLENGES OF THE FAST GROWING 3D SYSTEM INTEGRATION MARKET," YOLE DEVELOPPEMENT 2010.

[26] JUNG, M., SONG, T., WAN, Y., LEE, Y.-J., MOHAPATRA, D., WANG, H., TAYLOR, G., JARIWALA, D., PITCHUMANI, V., MORROW, P., WEBB, C., FISCHER, P., and LIM, S. K., "How to reduce power in 3d ic designs: A case study with opensparc t2 core," in *Custom Integrated Circuits Conference (CICC), 2013 IEEE*, pp. 1–4, Sept 2013.

[27] JUNG, M., SONG, T., WAN, Y., PENG, Y., and LIM, S. K., "On enhancing power benefits in 3D ICs: Block folding and bonding styles perspective," in *Design Automation Conference (DAC), 2014 51st ACM/EDAC/IEEE*, pp. 1–6, June 2014.

[28] KANG, U., CHUNG, H.-J., HEO, S., AHN, S.-H., LEE, H., CHA, S.-H., AHN, J., KWON, D., KIM, J. H., LEE, J.-W., JOO, H.-S., KIM, W.-S., KIM, H.-K., LEE, E.-M., KIM, S.-R., MA, K.-H., JANG, D.-H., KIM, N.-S., CHOI, M.-S., OH, S.-J., LEE, J.-B., JUNG, T.-K., YOO, J.-H., and KIM, C., "8Gb 3D DDR3 DRAM using through-silicon-via technology," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pp. 130–131,131a, Feb 2009.

[29] KARKLIN, K., BROZ, J., and MANN, B., "Bond Pad Damage Tutorial," in *IEEE Semiconductor Wafer Test Workshop*, 2010.

[30] KATTI, G., STUCCHI, M., DE MEYER, K., and DEHAENE, W., "Electrical modeling and characterization of through silicon via for three-dimensional ics," *Electron Devices, IEEE Transactions on*, vol. 57, pp. 256–262, Jan 2010.

[31] KATTI, G., STUCCHI, M., DE MEYER, K., and DEHAENE, W., "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *Electron Devices, IEEE Transactions on*, vol. 57, pp. 256–262, Jan 2010.

[32] KIM, D. H., ATHIKULWONGSE, K., HEALY, M., HOSSAIN, M., JUNG, M., KHOROSH, I., KUMAR, G., LEE, Y.-J., LEWIS, D., LIN, T.-W., LIU, C., PANTH, S., PATHAK, M., REN, M., SHEN, G., SONG, T., WOO, D. H., ZHAO, X., KIM, J., CHOI, H., LOH, G., LEE, H.-H., and LIM, S. K., "3D-MAPS: 3D Massively parallel processor with stacked memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 188–190, Feb 2012.

[33] KIM, D. H., ATHIKULWONGSE, K., and LIM, S. K., "A study of Through-Silicon-Via impact on the 3D stacked IC layout," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, pp. 674–680, Nov 2009.

[34] KIM, J., LEE, W., SHIM, Y., SHIM, J., KIM, K., PAK, J. S., and KIM, J., "Chip-Package Hierarchical Power Distribution Network Modeling and Analysis Based on a Segmentation Method," *Advanced Packaging, IEEE Transactions on*, vol. 33, pp. 647–659, Aug 2010.

[35] KIM, J., PAK, J. S., CHO, J., SONG, E., CHO, J., KIM, H., SONG, T., LEE, J., LEE, H., PARK, K., YANG, S., SUH, M.-S., BYUN, K.-Y., and KIM, J., "High-Frequency Scalable Electrical Model and Analysis of a Through Silicon Via (TSV)," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, pp. 181 –195, feb. 2011.

[36] KIM, J.-S., OH, C. S., LEE, H., LEE, D., HWANG, H.-R., HWANG, S., NA, B., MOON, J., KIM, J.-G., PARK, H., RYU, J.-W., PARK, K., KANG, S.-K., KIM, S.-Y., KIM, H., BANG, J.-M., CHO, H., JANG, M., HAN, C., LEE, J.-B., KYUNG, K., CHOI, J.-S., and JUN, Y.-H., "A 1.2V 12.8GB/s 2Gb mobile Wide-I/O DRAM with 4x128 I/Os using TSV-based stacking," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pp. 496–498, Feb 2011.

[37] KIM, W., GUPTA, M., WEI, G.-Y., and BROOKS, D., "System level analysis of fast, per-core DVFS using on-chip switching regulators," in *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*, pp. 123–134, Feb 2008.

[38] KIM, Y.-J. and ALLEN, M., "Integrated solenoid-type inductors for high frequency applications and their characteristics," in *Electronic Components amp; Technology Conference, 1998. 48th IEEE*, pp. 1247 –1252, may 1998.

[39] KUMAR, R. and KHATRI, S. P., "Crosstalk avoidance codes for 3D VLSI," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, pp. 1673–1678, March 2013.

[40] LEE, C.-K., ZHAN, C.-J., LAU, J., HUANG, Y.-J., FU, H.-C., HUANG, J.-H., HSIAO, Z.-C., CHEN, S.-W., HUANG, S.-Y., FAN, C.-W., LIN, Y.-M., KAO, K.-S., KO, C.-T., CHEN, T.-H., LO, R., and KAO, M., "Wafer bumping, assembly, and reliability assessment of ubumps with 5 um pads on 10 um pitch for 3D IC integration," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, pp. 636–640, May 2012.

[41] LEE, K., FUKUSHIMA, T., TANAKA, T., and KOYANAGI, M., "3D integration technology and reliability challenges," in *Electrical Design of Advanced Packaging and Systems Symposium (EDAPS), 2011 IEEE*, pp. 1–4, Dec 2011.

[42] LEE, Y.-J. and LIM, S. K., "Timing analysis and optimization for 3D stacked multi-core microprocessors," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, pp. 1 –7, nov. 2010.

[43] LEE, Y.-J. and LIM, S. K., "On GPU bus power reduction with 3D IC technologies," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–6, March 2014.

[44] LEE, Y.-J. and LIM, S. K., "On GPU bus power reduction with 3D IC technologies," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–6, March 2014.

[45] LIU, C., SONG, T., CHO, J., KIM, J., KIM, J., and LIM, S. K., "Full-chip TSV-to-TSV coupling analysis and optimization in 3D IC," in *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, pp. 783 –788, june 2011.

[46] LO, W.-C., CHEN, Y.-H., KO, J.-D., KUO, T.-Y., SHIH, Y.-C., and LU, S.-T., "An innovative chip-to-wafer and wafer-to-wafer stacking," in *Electronic Components and Technology Conference, 2006. Proceedings. 56th*, pp. 6 pp.–, 2006.

[47] MARTINI, M., "The long and tortuous path of EUV lithography to full production," 2014.

[48] MOTOYOSHI, M., TAKANOHASHI, J., FUKUSHIMA, T., ARAI, Y., and KOYANAGI, M., "Stacked SOI pixel detector using versatile fine pitch u-bump technology," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*, pp. 1–4, Jan 2012.

[49] MOTOYOSHI, M., TAKANOHASHI, J., FUKUSHIMA, T., ARAI, Y., and KOYANAGI, M., "Stacked SOI pixel detector using versatile fine pitch u-bump technology," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*, pp. 1–4, Jan 2012.

[50] MOTOYOSHI, M. and KOYANAGI, M., "3D-LSI technology for image sensor," *Journal of Instrumentation*, vol. 4, no. 03, p. P03009, 2009.

[51] MURUGESAN, M., KINO, H., HASHIGUCHI, A., MIYAZAKI, C., SHIMAMOTO, H., KOBAYASHI, H., FUKUSHIMA, T., TANAKA, T., and KOYANAGI, M., "High density 3D LSI technology using W/Cu hybrid TSVs," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 6.6.1–6.6.4, Dec 2011.

[52] NAWATHE, U., HASSAN, M., WARRINER, L., YEN, K., UPPUTURI, B., GREENHILL, D., KUMAR, A., and PARK, H., "An 8-core 64-thread 64b power-efficient sparc soc," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 108–590, Feb 2007.

[53] OHARA, Y., NORIKI, A., SAKUMA, K., LEE, K.-W., MURUGESAN, M., BEA, J., YAMADA, F., FUKUSHIMA, T., TANAKA, T., and KOYANAGI, M., "10 um fine pitch Cu/Sn micro-bumps for 3-D super-chip stack," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pp. 1–6, Sept 2009.

[54] OK, S.-H., BAE, K.-R., LIM, S. K., and MOON, B., "Design and analysis of 3d ic-based low power stereo matching processors," in *Low Power Electronics and Design, 2002. ISLPED '02. Proceedings of the 2002 International Symposium on*, pp. 15–20, Aug 2013.

[55] Opencore, [Online]. Available: http://opencores.org/.

[56] ORACLE, "OPENSPARC T2." [ONLINE]. AVAILABLE: HTTP://WWW.ORACLE.COM/.

[57] PACKAN, P., AKBAR, S., ARMSTRONG, M., BERGSTROM, D., BRAZIER, M., DESHPANDE, H., DEV, K., DING, G., GHANI, T., GOLONZKA, O., HAN, W., HE, J., HEUSSNER, R., JAMES, R., JOPLING, J., KENYON, C., LEE, S.-H., LIU, M., LODHA, S., MATTIS, B., MURTHY, A., NEIBERG, L., NEIRYNCK, J., PAE, S., PARKER, C., PIPES, L., SEBASTIAN, J., SEIPLE, J., SELL, B., SHARMA, A., SIVAKUMAR, S., SONG, B., ST.AMOUR, A., TONE, K., TROEGER, T., WEBER, C., ZHANG, K., LUO, Y., and NATARAJAN, S., "High performance 32nm logic technology featuring 2nd generation high-k + metal gate transistors," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, Dec 2009.

[58] PAN, S., CHANG, N., and ZHENG, J., "A new methodology for IC-package thermal co-analysis in 3D IC environment," in *Electrical Design of Advanced Packaging Systems Symposium (EDAPS), 2010 IEEE*, pp. 1–4, Dec 2010.

[59] PATHAK, M., LEE, Y.-J., MOON, T., and LIM, S.-K., "Through-silicon-via management during 3D physical design: When to add and how many?," in *Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on*, pp. 387–394, 2010.

[60] PATIL, N., LIN, A., ZHANG, J., WONG, H.-S., and MITRA, S., "Digital VLSI logic technology using Carbon Nanotube FETs: Frequently Asked Questions," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, pp. 304–309, July 2009.

[61] PAUL, C. R., *Analysis of multiconductor transmission lines*. Lexington, KY: John Wiley and Sons, 1994.

[62] PENG, L., LI, H., LIM, D. F., GAO, S., and TAN, C. S., "High-Density 3-D Interconnect of Cu-Cu Contacts With Enhanced Contact Resistance by Self-Assembled Monolayer (SAM) Passivation," *Electron Devices, IEEE Transactions on*, vol. 58, pp. 2500–2506, Aug 2011.

[63] http://ptm.asu.edu/.

[64] SALAH, K., RAGAI, H., and ISMAIL, Y., "A macro-modeling approach for through silicon via," in *EUROCON, 2013 IEEE*, pp. 1869–1872, July 2013.

[65] SAMSUNG STARTS MASS PRODUCING INDUSTRY'S FIRST 3D TSV TECHNOLOGY BASED DDR4 MODULES FOR ENTERPRISE SERVERS, HTTP://WWW.SAMSUNG.COM/.

[66] SCHROM, G., HAZUCHA, P., HAHN, J., GARDNER, D., BLOECHEL, B., DERMER, G., NARENDRA, S., KARNIK, T., and DE, V., "A 480-MHz, multi-phase interleaved buck DC-DC converter with hysteretic control," in *Power Electronics Specialists Conference, 2004. PESC 04. 2004 IEEE 35th Annual*, vol. 6, pp. 4702–4707 Vol.6, June 2004.

[67] SHI, J., POPOVIC, D., NETTLETON, N., SZE, T., DOUGLAS, D., THACKER, H., CUNNINGHAM, J., FURUTA, K., KOJIMA, R., HIROSE, K., and HWANG, K., "Direct chip powering and enhancement of proximity communication through Anisotropic Conductive adhesive chip-to-chip bonding," in *Electronic Components and Technology Conference (ECTC), 2010 Proceedings 60th*, pp. 363–368, June 2010.

[68] SMITH, L., ROY, T., and ANDERSON, R., "Power plane SPICE models for frequency and time domains," in *Electrical Performance of Electronic Packaging, 2000, IEEE Conference on.*, pp. 51–54, 2000.

[69] SONG, T., LIU, C., KIM, D. H., LIM, S. K., CHO, J., KIM, J., PAK, J. S., AHN, S., KIM, J., and YOON, K., "Analysis of TSV-to-TSV coupling with high-impedance termination in 3D ICs," in *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, pp. 1 –7, march 2011.

[70] SONG, T., LIU, C., PENG, Y., and LIM, S. K., "Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs," in *Design Automation Conference (DAC), 2013 50th ACM / EDAC / IEEE*, pp. 1–7, May 2013.

[71] SONG, T., NIEUWOUDT, A., YU, Y. S., and LIM, S. K., "Coupling capacitance in face-to-face (f2f) bonded 3d ics: Trends and implications," in *Electronic Components and Technology Conference (ECTC) , 2015 IEEE 65th*, pp. 529–536, May 2015.

[72] STURCKEN, N., O'SULLIVAN, E., WANG, N., HERGET, P., WEBB, B., ROMANKIW, L., PETRACCA, M., DAVIES, R., FONTANA, R., DECAD, G., KYMISSIS, I., PETERCHEV, A., CARLONI, L., GALLAGHER, W., and SHEPARD, K., "A 2.5D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer delivering 10.8A/mm2," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 400–402, Feb 2012.

[73] STURCKEN, N., PETRACCA, M., WARREN, S., CARLONI, L., PETERCHEV, A., and SHEPARD, K., "An integrated four-phase buck converter delivering 1A/mm2 with 700ps controller delay and network-on-chip load in 45-nm SOI," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, pp. 1–4, Sept 2011.

[74] SYNOPSYS, 32/28NM GENERIC LIBRARY. [ONLINE]. AVAILABLE: HTTP://WWW.SYNOPSYS.COM/.

[75] SYNOPSYS QUICKCAP NX, HTTP://WWW.SYNOPSYS.COM/.

[76] TECHINSIGHTS, "Package Analysis of the SK-Hynix High Bandwidth Memory," 2015.

[77] TYAGI, S., AUTH, C., BAI, P., CURELLO, G., DESHPANDE, H., GANNAVARAM, S., GOLONZKA, O., HEUSSNER, R., JAMES, R., KENYON, C., LEE, S.-H., LINDERT, N., LIU, M., NAGISETTY, R., NATARAJAN, S., PARKER, C., SEBASTIAN, J., SELL, B., SIVAKUMAR, S., ST AMOUR, A., and TONE, K., "An advanced low power, high performance, strained channel 65nm technology," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 245–247, Dec 2005.

[78] WANG, N., OSULLIVAN, E. J., HERGET, P., RAJENDRAN1, B., KRUPP, L. E., ROMANKIW1, L. T., WEBB, B. C., FONTANA, R., DUCH, E. A., JOSEPH, E. A., BROWN, S. L., HU, X., DECAD, G. M., STURCKEN, N., SHEPARD, K. L., and GALLAGHER, W. J., "Integrated on-chip inductors with electroplated magnetic yokes (invited)," *Journal of Applied Physics*, vol. 111, no. 7, p. 07E732, 2012.

[79] WESTE, N. and HARRIS, D., *CMOS VLSI Design: A Circuits and Systems Perspective*. USA: Addison-Wesley Publishing Company, 4th ed., 2010.

[80] WIBBEN, J. and HARJANI, R., "A High Efficiency DC-DC Converter Using 2nH On-Chip Inductors," in *VLSI Circuits, 2007 IEEE Symposium on*, pp. 22 –23, june 2007.

[81] XILINX, "Xcell Journal, Issue 74," 2011.

[82] XU, C., JIANG, L., KOLLURI, S., RUBIN, B., DEUTSCH, A., SMITH, H., and BANERJEE, K., "Fast 3-D thermal analysis of complex interconnect structures using electrical modeling and simulation methodologies," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, pp. 658–665, Nov 2009.

[83] YOON, K., KIM, G., LEE, W., SONG, T., LEE, J., LEE, H., PARK, K., and KIM, J., "Modeling and analysis of coupling between TSVs, metal, and RDL interconnects in TSV-based 3D IC with silicon interposer," in *Electronics Packaging Technology Conference, 2009. EPTC '09. 11th*, pp. 702–706, Dec 2009.